TORONTO PUBLIC LIBRARY
~~37131 056 339 864~~
BF Barbara Frum

A Guided Tour of Mathematical Methods for the Physical Sciences

Second Edition

Roel Snieder

CAMBRIDGE

A GUIDED TOUR OF MATHEMATICAL METHODS

For the Physical Sciences

Second Edition

Mathematical methods are essential tools for all physical scientists. This second edition of *A Guided Tour of Mathematical Methods* provides a comprehensive tour of the mathematical knowledge and techniques that are needed by students in this area. In contrast to more traditional textbooks, all the material is presented in the form of problems. Within these problems the basic mathematical theory and its physical applications are well integrated. The mathematical insights that the student acquires are therefore driven by their physical insight. Topics that are covered include vector calculus, linear algebra, Fourier analysis, scale analysis, complex integration, Green's functions, normal modes, tensor calculus, and perturbation theory. The second edition contains new chapters on dimensional analysis, variational calculus, and the asymptotic evaluation of integrals. This book can be used by undergraduates and lower-level graduate students in the physical sciences. It can serve as a stand-alone text, or as a source of problems and examples to complement other textbooks.

ROEL SNIEDER holds the Keck Foundation Endowed Chair of Basic Exploration Science at the Colorado School of Mines. He received his Masters degree in Geophysical Fluid Dynamics from Princeton University in 1984 and in 1987 a Ph.D. in seismology from Utrecht University. In 1993 he was appointed as professor of seismology at Utrecht University, where from 1997 to 2000 he served as Dean of the Faculty of Earth Sciences. In 1997 he was a visiting professor at the Center for Wave Phenomena at the Colorado School of Mines. His research focuses on wave propagation and inverse problems. He has served on the editorial boards of Geophysical Journal International, Inverse Problems, and Reviews of Geophysics. In 2000 he was elected as Fellow of the American Geophysical Union for important contributions to geophysical inverse theory, seisic tomography, and the theory of surface waves.

A GUIDED TOUR OF MATHEMATICAL METHODS

For the Physical Sciences

Second Edition

ROEL SNIEDER
Department of Geophysics and Center for Wave Phenomena, Colorado School of Mines

CAMBRIDGE
UNIVERSITY PRESS

PUBLISHED BY THE PRESS SYNDICATE OF THE UNIVERSITY OF CAMBRIDGE
The Pitt Building, Trumpington Street, Cambridge, United Kingdom

CAMBRIDGE UNIVERSITY PRESS
The Edinburgh Building, Cambridge CB2 2RU, UK
40 West 20th Street, New York, NY 10011–4211, USA
477 Williamstown Road, Port Melbourne, VIC 3207, Australia
Ruiz de Alarcón 13, 28014 Madrid, Spain
Dock House, The Waterfront, Cape Town 8001, South Africa

http://www.cambridge.org

© R. Snieder 2004

This book is in copyright. Subject to statutory exception
and to the provisions of relevant collective licensing agreements,
no reproduction of any part may take place without
the written permission of Cambridge University Press.

First published 2004

Printed in the United Kingdom at the University Press, Cambridge

Typeface Times 11/14 pt. *System* $\LaTeX\,2_\varepsilon$ [TB]

A catalog record for this book is available from the British Library

Library of Congress Cataloging in Publication data

Snieder, Roel, 1958–
A guided tour of mathematical methods for the physical sciences/Roel Snieder – 2nd ed.
p. cm
Includes bibliographical references and index.
ISBN 0 521 83492 9 (hardback)
1. Mathematical analysis. 2. Physical sciences – Mathematics. 3. Mathematical physics. I. Title.
QA300.S794 2004
515–dc22 2004040783

ISBN 0 521 83492 9 hardback

The publisher has used its best endeavors to ensure that the URLs for external websites referred to in this book are correct and active at the time of going to press. However, the publisher has no responsibility for the websites and can make no guarantee that a site will remain live or that the content is or will remain appropriate.

To Idske, Hylke, Hidde and Julia

Contents

Preface to Second Edition *page* xiii
Acknowledgements xiv
1 Introduction 1
2 Dimensional analysis 3
2.1 Two rules for physical analysis 3
2.2 A trick for finding mistakes 6
2.3 Buckingham pi theorem 7
2.4 Lift of a wing 11
2.5 Scaling relations 12
2.6 Dependence of pipe flow on the radius of the pipe 13
3 Power series 16
3.1 Taylor series 16
3.2 Growth of the Earth by cosmic dust 22
3.3 Bouncing ball 24
3.4 Reflection and transmission by a stack of layers 27
4 Spherical and cylindrical coordinates 31
4.1 Introducing spherical coordinates 31
4.2 Changing coordinate systems 35
4.3 Acceleration in spherical coordinates 37
4.4 Volume integration in spherical coordinates 40
4.5 Cylindrical coordinates 43
5 Gradient 46
5.1 Properties of the gradient vector 46
5.2 Pressure force 50
5.3 Differentiation and integration 53
5.4 Newton's law from energy conservation 55
5.5 Total and partial time derivatives 57
5.6 Gradient in spherical coordinates 61

6 Divergence of a vector field 64
6.1 Flux of a vector field 64
6.2 Introduction of the divergence 66
6.3 Sources and sinks 69
6.4 Divergence in cylindrical coordinates 71
6.5 Is life possible in a five-dimensional world? 73
7 Curl of a vector field 78
7.1 Introduction of the curl 78
7.2 What is the curl of the vector field? 80
7.3 First source of vorticity: rigid rotation 81
7.4 Second source of vorticity: shear 83
7.5 Magnetic field induced by a straight current 85
7.6 Spherical coordinates and cylindrical coordinates 86
8 Theorem of Gauss 88
8.1 Statement of Gauss's law 88
8.2 Gravitational field of a spherically symmetric mass 89
8.3 Representation theorem for acoustic waves 91
8.4 Flowing probability 93
9 Theorem of Stokes 97
9.1 Statement of Stokes's law 97
9.2 Stokes's theorem from the theorem of Gauss 100
9.3 Magnetic field of a current in a straight wire 102
9.4 Magnetic induction and Lenz's law 103
9.5 Aharonov–Bohm effect 104
9.6 Wingtips vortices 108
10 Laplacian 113
10.1 Curvature of a function 113
10.2 Shortest distance between two points 117
10.3 Shape of a soap film 120
10.4 Sources of curvature 124
10.5 Instability of matter 126
10.6 Where does lightning start? 128
10.7 Laplacian in spherical and cylindrical coordinates 129
10.8 Averaging integrals for harmonic functions 130
11 Conservation laws 133
11.1 General form of conservation laws 133
11.2 Continuity equation 135
11.3 Conservation of momentum and energy 136
11.4 Heat equation 140
11.5 Explosion of a nuclear bomb 145

11.6 Viscosity and the Navier–Stokes equation 147
11.7 Quantum mechanics and hydrodynamics 150
12 Scale analysis 153
12.1 Vortex in a bathtub 154
12.2 Three ways to estimate a derivative 156
12.3 Advective terms in the equation of motion 159
12.4 Geometric ray theory 162
12.5 Is the Earth's mantle convecting? 167
12.6 Making an equation dimensionless 169
13 Linear algebra 173
13.1 Projections and the completeness relation 173
13.2 Projection on vectors that are not orthogonal 177
13.3 Coriolis force and centrifugal force 179
13.4 Eigenvalue decomposition of a square matrix 184
13.5 Computing a function of a matrix 187
13.6 Normal modes of a vibrating system 189
13.7 Singular value decomposition 192
13.8 Householder transformation 197
14 Dirac delta function 202
14.1 Introduction of the delta function 202
14.2 Properties of the delta function 206
14.3 Delta function of a function 208
14.4 Delta function in more dimensions 210
14.5 Delta function on the sphere 210
14.6 Self energy of the electron 212
15 Fourier analysis 217
15.1 Real Fourier series on a finite interval 217
15.2 Complex Fourier series on a finite interval 221
15.3 Fourier transform on an infinite interval 223
15.4 Fourier transform and the delta function 224
15.5 Changing the sign and scale factor 225
15.6 Convolution and correlation of two signals 228
15.7 Linear filters and the convolution theorem 231
15.8 Dereverberation filter 234
15.9 Design of frequency filters 238
15.10 Linear filters and linear algebra 240
16 Analytic functions 245
16.1 Theorem of Cauchy–Riemann 245
16.2 Electric potential 249
16.3 Fluid flow and analytic functions 251

17 Complex integration 254
17.1 Nonanalytic functions 254
17.2 Residue theorem 255
17.3 Solving integrals without knowing the primitive function 259
17.4 Response of a particle in syrup 262
18 Green's functions: principles 267
18.1 Girl on a swing 267
18.2 You have seen Green's functions before! 272
18.3 Green's functions as impulse response 273
18.4 Green's functions for a general problem 276
18.5 Radiogenic heating and the Earth's temperature 279
18.6 Nonlinear systems and the Green's functions 284
19 Green's functions: examples 288
19.1 Heat equation in *N* dimensions 288
19.2 Schrödinger equation with an impulsive source 292
19.3 Helmholtz equation in one, two, and three dimensions 296
19.4 Wave equation in one, two, and three dimensions 302
19.5 If I can hear you, you can hear me 308
20 Normal modes 311
20.1 Normal modes of a string 312
20.2 Normal modes of a drum 314
20.3 Normal modes of a sphere 317
20.4 Normal modes of orthogonality relations 323
20.5 Bessel functions behave as decaying cosines 327
20.6 Legendre functions behave as decaying cosines 330
20.7 Normal modes and the Green's function 334
20.8 Guided waves in a low-velocity channel 340
20.9 Leaky modes 344
20.10 Radiation damping 348
21 Potential theory 353
21.1 Green's function of the gravitational potential 354
21.2 Upward continuation in a flat geometry 356
21.3 Upward continuation in a flat geometry in three dimensions 359
21.4 Gravity field of the Earth 361
21.5 Dipoles, quadrupoles, and general relativity 365
21.6 Multipole expansion 369
21.7 Quadrupole field of the Earth 374
21.8 Fifth force 377
22 Cartesian tensors 379
22.1 Coordinate transforms 379

22.2 Unitary matrices 382
22.3 Shear or dilatation? 385
22.4 Summation convention 389
22.5 Matrices and coordinate transforms 391
22.6 Definition of a tensor 393
22.7 Not every vector is a tensor 396
22.8 Products of tensors 398
22.9 Deformation and rotation again 401
22.10 Stress tensor 403
22.11 Why pressure in a fluid is isotropic 406
22.12 Special relativity 408
23 Perturbation theory 412
23.1 Regular perturbation theory 413
23.2 Born approximation 417
23.3 Linear travel time tomography 421
23.4 Limits on perturbation theory 424
23.5 WKB approximation 427
23.6 Need for consistency 431
23.7 Singular perturbation theory 433
24 Asymptotic evaluation of integrals 437
24.1 Simplest tricks 437
24.2 What does $n!$ have to do with e and $\sqrt{\pi}$? 441
24.3 Method of steepest descent 445
24.4 Group velocity and the method of stationary phase 450
24.5 Asymptotic behavior of the Bessel function $J_0(x)$ 453
24.6 Image source 456
25 Variational calculus 461
25.1 Designing a can 461
25.2 Why are cans round? 463
25.3 Shortest distance between two points 465
25.4 The great-circle 468
25.5 Euler–Lagrange equation 472
25.6 Lagrangian formulation of classical mechanics 476
25.7 Rays are curves of stationary travel time 478
25.8 Lagrange multipliers 481
25.9 Designing a can with an optimal shape 485
25.10 The chain line 487
26 Epilogue, on power and knowledge 492
References 494
Index 500

Preface to Second Edition

The updates and changes from the earlier version of the book have to a large extent been driven by the comments of readers and reviewers. The second edition has been extended with Chapters 2, 24, and 25 that cover dimensional analysis, the asymptotic evaluation of integrals, and variational calculus, respectively. In a few places, new material has been inserted, such as Section 19.5 that covers the reciprocity of wave propagation. A number of teachers and students remarked that the level of difficulty of the problems in the first edition was highly variable. The problems in the second edition contain more hints and advice to make these problems more tractable.

Acknowledgements

This book resulted from two courses on mathematical physics that were taught at Utrecht University. The remarks, corrections, and the encouragement of a large number of students have been very important in its development. It is impossible to thank all the students, but I especially want to thank the feedback from Jojanneke van den Berg, Jehudi Blom, Sterre Dortland, Thomas Geenen, Wiebe van Driel, Luuk van Gerven, Noor Hogeweg, and Frederiek Siegenbeek. In their role as teaching assistants, Dirk Kraaipoel and Jesper Spetzler have helped greatly in improving this book. Huub Douma has spent numerous hours at sea correcting earlier drafts. A number of colleagues have helped me very much with their comments; I especially want to mention Freeman Gilbert, Alexander Kaufman, Antoine Khater, Ken Larner, and Jeannot Trampert. Many of the figures were drafted by Barbara McLenon, who I thank for her support. The help of Joop Hoofd, Everhard Muyzert, and John Stockwell, who patiently coped with my computer illiteracy, allowed me to prepare this book electronically. Terry Young greatly helped me to make the index of the second edition. The support and advice of Adam Black, Eoin O'Sullivan, Jayne Aldhouse, Maureen Storey, Joseph Bottrill, Keith Westmoreland and Simon Capelin of Cambridge University Press has been very helpful and stimulating during the preparation of this work. Lastly, I want to thank everybody who helped me in numerous ways to make writing this book a joy.

1

Introduction

The topic of this book is the application of mathematics to physical problems. Mathematics and physics are often taught separately. Despite the fact that education in physics relies on mathematics, it turns out that students consider mathematics to be disjoint from physics. Although this point of view may strictly be correct, it reflects an erroneous opinion when it concerns an education in the sciences. The reason for this is that mathematics is the *only* language at our disposal for quantifying physical processes. One cannot learn a language by just studying a textbook. In order to truly learn how to use a language one has to go abroad and start using that language. By the same token one cannot learn how to use mathematics in the physical sciences by just studying textbooks or attending lectures; the only way to achieve this is to venture into the unknown and apply mathematics to physical problems.

It is the goal of this book to do exactly that; problems are presented in order to apply mathematical techniques and knowledge to physical concepts. These examples are not presented as well-developed theory. Instead, they are presented as a number of problems that elucidate the issues that are at stake. In this sense this book offers a guided tour: material for learning is presented but true learning will only take place by active exploration. In this process, the interplay of mathematics and physics is essential; mathematics is the natural language for physics while physical insight allows for a better understanding of the mathematics that is presented.

How can you use this book most efficiently?

Since this book is written as a set of problems you may frequently want to consult other material as well to refresh or deepen your understanding of material. In many places we refer to the book of Boas [19]. In addition, the books of Butkov [24], Riley *et al.* [87] and Arfken [5] on mathematical physics are excellent.

In addition to books, colleagues in either the same field or other fields can be a great source of knowledge and understanding. Therefore, do not hesitate to work together with others on these problems if you are in the fortunate position to do so. This may not only make the work more enjoyable, it may also help you in getting "unstuck" at difficult moments and the different viewpoints of others may help to deepen yours.

For who is this book written?

This book is set up with the goal of obtaining a good working knowledge of mathematical physics that is needed for students in physics or geophysics. A certain basic knowledge of calculus and linear algebra is required to digest the material presented here. For this reason, this book is meant for upper-level undergraduate students or lower-level graduate students, depending on the background and skill of the student. In addition, teachers can use this book as a source of examples and illustrations to enrich their courses.

This book is evolving

This book will be improved regularly by adding new material, correcting errors and making the text clearer. The feedback of both teachers and students who use this material is vital in improving this text, please send your remarks to:

Roel Snieder

Dept of Geophysics

Colorado School of Mines
Golden CO 80401
USA

telephone: +1-303-273.3456
fax: +1-303-273.3478
email: rsnieder@mines.edu

Errata can be found at the following website: www.mines.edu/~rsnieder/Errata.html.

2

Dimensional analysis

The material of this chapter is usually not covered in a book on mathematics. The field of mathematics deals with numbers and numerical relationships. It does not matter what these numbers are; they may account for physical properties of a system, but they may equally well be numbers that are not related to anything physical. Consider the expression $g = df/dt$. From a mathematical point of view these functions can be anything, as long as g is the derivative of f. The situation is different in physics. When $f(t)$ is the position of a particle, and t denotes time, then $g(t)$ is a velocity. This relation fixes the physical dimension of $g(t)$. In mathematical physics, the physical dimension of variables imposes constraints on the relation between these variables. In this chapter we explore these constraints. In Section 2.2 we show that this provides a powerful technique for spotting errors in equations. In the remainder of this chapter we show how the physical dimensions of the variables that govern a problem can be used to find physical laws. Surprisingly, while most engineers learn about dimensional analysis, this topic is not covered explicitly in many science curricula.

2.1 Two rules for physical dimensions

In physics every physical parameter is associated with a physical dimension. The value of each parameter is measured with a certain physical unit. For example, when I measure how long a table is, the result of this measurement has dimension "length". This length is measured in a certain unit, that may be meters, inches, furlongs, or whatever length unit I prefer to use. The result of this measurement can be written as

$$l = 3 \text{ m}. \tag{2.1}$$

The variable l has the physical dimension of length, in this chapter we write this as

$$l \sim [L]. \tag{2.2}$$

The square brackets are used in this chapter to indicate a physical dimension. The capital letter L denotes length, T denotes time, and M denotes mass. Other physical dimensions include electric charge and temperature. When dealing with physical dimensions two rules are useful. The first rule is:

Rule 1 When two variables are added, subtracted, or set equal to each other, they must have the same physical dimension.

In order to see the logic of this rule we consider the following example. Suppose we have an object with a length of 1 meter and a time interval of one second. This means that

$$\begin{aligned} l &= 1 \text{ m}, \\ t &= 1 \text{ s}. \end{aligned} \tag{2.3}$$

Since both variables have the same numerical value, we might be tempted to declare that

$$l = t. \tag{2.4}$$

It is, however, important to realize that the physical units that we use are arbitrary. Suppose, for example, that we had measured the length in feet rather than meters. In that case the measurements (2.3) would be given by

$$\begin{aligned} l &= 3 \text{ ft}, \\ t &= 1 \text{ s}. \end{aligned} \tag{2.5}$$

Now the numerical value of the same length measurement is different! Since the choice of the physical units is arbitrary, we can scale the relation between variables of different physical dimensions in an arbitrary way. For this reason these variables cannot be equal to each other. This implies that they cannot be added or subtracted either.

The first rule implies the following rule.

Rule 2 Mathematical functions can act on dimensionless numbers only.

To see this, let us consider as an example the function $f(\xi) = e^{\xi}$. Using a Taylor expansion, this function can be written as:

$$f(\xi) = 1 + \xi + \frac{1}{2}\xi^2 + \cdots \tag{2.6}$$

According to rule 1 the different terms in this expression must have the same physical dimension. The first term (the number 1) is dimensionless, hence all the other terms in the series must be dimensionless. This means that ξ must be a dimensionless number as well. This argument can be used for any function $f(\xi)$ whose Taylor expansion contains different powers of ξ. Note that the argument would not hold for a function such as $f(\xi) = \xi^2$ that contains only one power of ξ. To please the purists, rule 2 could easily be reformulated to exclude these special cases.

These rules have several applications in mathematical physics. Suppose we want to find the physical dimension of a force, as expressed in the basic dimensions mass, length, and time. The only thing we need to do is take one equation that contains a force. In this case Newton's law $F = ma$ comes to mind. The mass m has physical dimension $[M]$, while the acceleration has dimension $[L/T^2]$. Rule 1 implies that force has the physical dimension $[ML/T^2]$.

Problem a The force F in a linear spring is related to the extension x of the spring by the relation $F = -kx$. Show that the spring constant k has dimension $[M/T^2]$.

Problem b The angular momentum $\mathbf{L}$ of a particle with momentum $\mathbf{p}$ at position $\mathbf{r}$ is given by

$$\mathbf{L} = \mathbf{r} \times \mathbf{p}, \tag{2.7}$$

where × denotes the cross-product of two vectors. Show that angular momentum has the dimension $[ML^2/T]$.

Problem c A plane wave is given by the expression

$$u(\mathbf{r}, t) = e^{i(\mathbf{k}\cdot\mathbf{r}-\omega t)}, \tag{2.8}$$

where $\mathbf{r}$ is the position vector and t denotes time. Show that $\mathbf{k} \sim [L^{-1}]$ and $\omega \sim [T^{-1}]$.

In quantum mechanics the behavior of a particle is characterized by a wave equation, that is called the Schrödinger equation. In one space dimension this equation is given by

$\frac{ML^2}{T^2} \cdot \psi = \hbar \frac{\psi}{T}$

$$i\hbar\frac{\partial\psi}{\partial t} = -\frac{\hbar^2}{2m}\frac{\partial^2\psi}{\partial x^2} + V(x)\psi, \tag{2.9}$$

where x denotes the position, t denotes the time, m the mass of the particle, and $V(x)$ the potential energy of the particle. At this point it is not clear what the wave

$\hbar\frac{\psi}{T} = \frac{\hbar^2\psi}{ML^2}$

function $\psi(x, t)$ is, and how this equation should be interpreted. The meaning of the symbol $\hbar$ is not yet defined. We can, however, determine the physical dimension of $\hbar$ without knowing the meaning of this variable.

Problem d Compare the physical dimensions of the left-hand side of (2.9) with the first term on the right-hand side and show that the variable $\hbar$ has the physical dimension angular momentum. You can use problem b in showing this.

2.2 A trick for finding mistakes

The requirement that all terms in an equation have the same physical dimension is an important tool for spotting mistakes. Cipra [26] gives many useful tips for spotting errors in his delightful book "Misteakes [sic]... and how to find them before the teacher does." As an example of using dimensional analysis for spotting mistakes, we consider the erroneous equation

$$E = mc^3, \qquad \frac{ML^3}{T^3} \tag{2.10}$$

where E denotes energy, m denotes mass, and c is the speed of light. Let us first find the physical dimension of energy. The work done by a force $\mathbf{F}$ over a displacement $d\mathbf{r}$ is given by $dE = \mathbf{F} \cdot d\mathbf{r}$. We showed in Section 2.1 that force has the dimension $[ML/T^2]$. This means that energy has the dimension $[ML^2/T^2]$. The speed of light in the right-hand side of expression (2.10) has dimension $[L/T]$, which means that the right-hand side has physical dimension $[ML^3/T^3]$. This is not an energy, which has dimension $[ML^2/T^2]$. Therefore expression (2.10) is wrong.

Problem a Now that we have determined that expression (2.10) is incorrect we can use the requirement that the dimensions of the different terms must match to guess how to set it right. Show that the right-hand side must be divided by a velocity to match the dimensions.

It is not clear that the right-hand side must be divided by the speed of light to give the correct expression $E = mc^2$. Dimensional analysis tells us only that it must be divided by something with the dimension of velocity. For all we know, it could be the speed at which the average snail moves.

Problem b Is the following equation dimensionally correct?

$$(\mathbf{v} \cdot \nabla \mathbf{v}) = -\nabla p. \tag{2.11}$$

In this expression $\mathbf{v}$ is the velocity of fluid flow, p is the pressure, and ∇ is the gradient vector (which essentially is a derivative with respect to the space

$\frac{L}{T^2}$ $\qquad$ $\frac{ML}{T^2} \cdot \frac{1}{L^2}$

coordinates). You can use that pressure has the dimension is force per unit area.

Problem c Answer the same question for the expression that relates the particle velocity v to the pressure p in an acoustic medium:

$$v = \frac{p}{\rho c} \qquad \frac{L}{T} = \frac{\cancel{M}\cancel{L}}{T^{\cancel{2}}\cancel{L^2}}\,\frac{\cancel{L^3}}{\cancel{M}}\,\frac{\cancel{T}}{\cancel{L}} \tag{2.12}$$

Here ρ is the mass density and c is velocity of propagation of acoustic waves.

Problem d In quantum mechanics, the energy E of the harmonic oscillator is given by

$$E_n = \hbar\omega^2 (n + 1/2), \qquad M\frac{L^2}{T^2} = \frac{ML^2}{T}\,T \tag{2.13}$$

where ω is a frequency, n is a dimensionless integer, and $\hbar$ is Planck's constant divided by 2π as introduced in problem d of the previous section. Verify if this expression is dimensionally correct.

In general it is a good idea to carry out a dimensional analysis while working in mathematical physics because this may help in finding the mistakes that we all make while doing derivations. It takes a little while to become familiar with the dimensions of properties that are used most often, but this is an investment that pays off in the long run.

2.3 Buckingham pi theorem

In this section we introduce the Buckingham pi theorem. This theorem can be used to find the relation between physical parameters based on dimensional arguments. As an example, let us consider a ball shown in Figure 2.1 with mass m that is dropped from a height h. We want to find the velocity with which it strikes the ground. The potential energy of the ball before it is dropped is mgh, where g is the acceleration of gravity. This energy is converted into kinetic energy $\frac{1}{2}mv^2$ as it strikes the ground. Equating these quantities and solving for the velocity gives:

$$v = \sqrt{2gh}. \tag{2.14}$$

Now let us suppose we did not know about classical mechanics. In that case, dimensional analysis could be used to guess relation (2.14). We know that the velocity is some function of the acceleration of gravity, the initial height, and the mass of the particle: $v = f(g, h, m)$. The physical dimensions of these properties

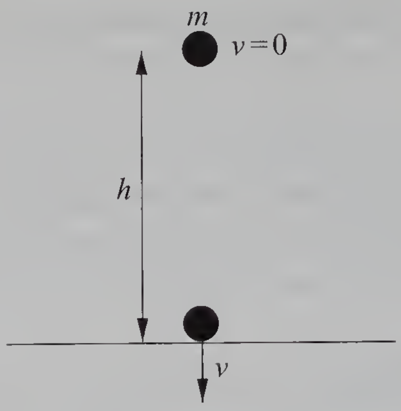

Fig. 2.1 Definition of the variables for a falling ball.

are given by

$$v \sim [L/T], \quad g \sim [L/T^2], \quad h \sim [L], \quad m \sim [M]. \tag{2.15}$$

Let us consider the dimension mass first. The dimension mass enters only the variable m. We cannot combine the variable m with the parameters g and h in any way to arrive at a quantity that is independent of mass. Therefore, the velocity does not depend on the mass of the particle. Next we consider the dimension time. The velocity depends on time as $\left[T^{-1}\right]$, the acceleration of gravity as $\left[T^{-2}\right]$, and h is independent of time. This means that we can match the dimension time only when

$$v \sim \sqrt{g}. \tag{2.16}$$

In this expression the left-hand side depends of the length as $[L]$, while the right-hand side varies with length as $[L^{1/2}]$. We have, however, not used the height h yet. The dimension length can be made to match if we multiply the right-hand side with $h^{1/2}$. This means that the only combination of g and h that gives a velocity is given by

$$v \sim \sqrt{gh}. \tag{2.17}$$

This result agrees with expression (2.14), which was derived using classical mechanics. Note that in order to arrive at expression (2.17) we used only dimensional arguments, and did not need to have any knowledge from classical mechanics other than that the velocity depends only on g and h. The dimensional analysis that led to expression (2.17), however, does not tell us what is the proportionality constant in that expression. The reason is that a proportionality constant is dimensionless, and can therefore not be found by dimensional analysis.

The treatment given here may appear to be cumbersome. This analysis, however, can be carried out in a systematic fashion using the *Buckingham pi theorem* [23] which states the following:

Buckingham pi theorem If a problem contains N variables that depend on P physical dimensions, then there are $N - P$ dimensionless numbers that describe the physics of the problem.

The original paper of Buckingham is very clear, but as we will see at the end of this section, this theorem is not fool-proof. Let us first apply the theorem to the problem of the falling ball. We have four variables: v, g, h, and m, so that $N = 4$. These variables depend on the physical dimensions $[M]$, $[L]$, and $[T]$, hence $P = 3$. According to the Buckingham pi theorem, $N - P = 1$ dimensionless number characterizes the problem. We want to express the velocity in the other parameters; hence we seek a dimensionless number of the form

$$vg^{\alpha}h^{\beta}m^{\gamma} \sim [1], \tag{2.18}$$

where the notation in the right-hand side means that it is dimensionless. Let us seek the exponents α, β, and γ that make the left-hand side dimensionless. Inserting the dimensions of the different variables then gives the following dimensions

$$\left[\frac{L}{T}\right]\left[\frac{L^{\alpha}}{T^{2\alpha}}\right]\left[L^{\beta}\right]\left[M^{\gamma}\right] \sim [1]. \tag{2.19}$$

The left-hand side depends on length as $[L^{1+\alpha+\beta}]$. The left-hand side can only be independent of length when the exponent is equal to zero. Applying the same reasoning to each of the dimensions length, time, and mass, then gives

$$\begin{array}{ll} \text{dimension } [L]: & 1 + \alpha + \beta = 0, \\ \text{dimension } [T]: & -1 - 2\alpha = 0, \\ \text{dimension } [M]: & \gamma = 0. \end{array} \tag{2.20}$$

This constitutes a system of three equations with three unknowns.

Problem a Show that the solution of this system is given by

$$\alpha = \beta = -\frac{1}{2}, \qquad \gamma = 0. \tag{2.21}$$

Inserting these values into expression (2.18) shows that the combination $vg^{-1/2}h^{-1/2}$ is dimensionless. This implies that

$$v = C\sqrt{gh}, \tag{2.22}$$

where C is the one dimensionless number in the problem as dictated by the Buckingham pi theorem.

The approach taken here is systematic. In his original paper [23], Buckingham applied this treatment to a number of problems: the thrust provided by the screw of a ship, the energy density of the electromagnetic field, the relation between the mass and radius of the electron, the radiation of an accelerated electron, and heat conduction.

There is, however, a catch that we introduce with an example. When air (or water) has a stably stratified mass–density structure, it can support oscillations where the restoring force is determined by the density gradient in the air. These oscillations occur with the *Brunt-Väisälä* frequency ω_B given by [50, 82]:

$$\omega_B = \sqrt{\frac{g}{\theta}\frac{d\theta}{dz}}. \tag{2.23}$$

In this expression, g is the acceleration of gravity, z is height, and θ is potential temperature (a measure of the thermal structure of the atmosphere).

Problem b Verify that this expression is dimensionally correct.

Problem c Check that this expression is also dimensionally correct when θ is replaced by the air pressure p, or the mass density ρ.

The result of problem c indicates that the potential temperature θ can be replaced by any physical parameter, and expression (2.23) is still dimensionally correct. This means that a dimensional analysis alone can never be used to prove that θ should be the potential temperature. In order to show this we need to know more of the physics of the problem.

Another limitation of the Buckingham pi theorem as formulated in its original form is that the theorem assumes that physical parameters need to be multiplied or divided to form dimensionless numbers; see equation (3) of reference [23]. The derivative of one variable with respect to another, however, has the same dimension as the ratio of these variables. Consider for example a problem where dimensional analysis shows that the variable of interest depends on the ratio of the acceleration of gravity and the height: g/h. The derivative of g with height dg/dz has the same physical dimension as g/h. Therefore, a dimensional analysis alone cannot completely describe the physics of the problem. Nevertheless, as we will see in the following section, it may provide valuable insights.

2.4 Lift of a wing

In this section we study the lift of a wing. Since in stationary flight the lift provided by a wing is equal to the weight of the aircraft or bird that is carried by the wing, we denote the lift of the wing with the symbol W. Since the lift is a force, this quantity has the dimension force: $W \sim [F] = [ML/T^2]$. The lift depends on the mass density ρ of the air, the velocity v of the airflow, and the surface area S of the wing.

Problem a Show that $\rho \sim [M/L^3]$, $v \sim [L/T]$, and $S \sim [L^2]$.

Problem b Count the number of variables and number of physical dimensions to show that in the jargon of the Buckingham pi theorem $N = 4$ and $P = 3$.

This means that there is $N - P = 1$ dimensionless number that characterizes the lift of the wing. We want to express the lift W in the other parameters, therefore we seek a dimensionless number of the form

$$W \rho^{\alpha} v^{\beta} S^{\gamma} \sim [1]. \tag{2.24}$$

$$\frac{ML}{T^2}\left(\frac{M}{L^3}\right)^{\alpha}\left(\frac{L}{T}\right)^{\beta}\left(L^2\right)^{\gamma} = C$$

Problem c Show that the requirement that the left-hand side does not depend on mass, length, and time, respectively, leads to the following linear equations:

$$\begin{aligned} 1 + \alpha &= 0, \\ 1 - 3\alpha + \beta + 2\gamma &= 0, \\ 2 + \beta &= 0. \end{aligned} \tag{2.25}$$

Problem d Solve this system to derive that

$$\alpha = \gamma = -1, \qquad \beta = -2. \tag{2.26}$$

Inserting this result in expression (2.24) implies that $W/(\rho v^2 S)$ is a dimensionless number. When this constant is denoted by C_L, this means that the lift is given by

$$W = C_L \rho v^2 S. \tag{2.27}$$

The coefficient C_L is called the *lift coefficient* [55]. This coefficient depends on the shape of the wing, and on the angle of attack. (This is a measure of the orientation of the wing to the airflow.) Let us think about the solution (2.27) for a moment. This expression states that the lift is proportional to the surface area; this makes sense: a larger wing produces more lift. The lift depends on the square of the velocity. It stands to reason that a larger flow velocity gives a larger lift, but that the lift increases quadratically with the velocity is not easy to see. Lastly, the lift is proportional to the mass density of the air: for a given velocity heavier air provides a larger lift because the airflow deflected by the wing has a larger momentum.

This has implications for the design of airports. For example, the airport of Denver is located at an elevation of about 1600 meters. This high elevation, in combination with the warm temperatures in summertime, leads to a relatively small mass density of the air. Since the surface area of the wings of aircraft is fixed by their design, the relatively small mass density can be compensated by a larger take-off velocity v only. In order to achieve this large take-off velocity, aircraft need a longer runway to accelerate to the required take-off velocity. For this reason, the airport in Denver has extra long runways. All these conclusions follow from dimensional analysis only!

2.5 Scaling relations

We can take the dimensional analysis of the previous section even a step further. Suppose we consider different flying objects, and that each object is characterized by a linear dimension l.

Problem a Use dimensional arguments to show that the volume V scales with the size as $V \sim l^3$, and that the surface area scales as $S \sim l^2$. (The volume V should not be confused with the velocity v.)

Problem b Show that this implies that

$$S \sim V^{2/3}. \tag{2.28}$$

The mass of the flying object is proportional to its mass density ρ_f by the relation $m = \rho_f V$. The lift required to support this mass is given by

$$W = g\rho_f V. \tag{2.29}$$

Problem c Insert the relations (2.28) and (2.29) into expression (2.27) to show that

$$g\rho_f V^{1/3} = C_L \rho v^2. \tag{2.30}$$

Problem d Solve this expression for V, and insert this result into expression (2.29) to derive the following relation between the lift and the velocity:

$$W = \frac{C_L^3 \rho^3}{g^2 \rho_f^2} v^6. \tag{2.31}$$

This expression predicts that the lift varies with the velocity to the sixth power. Figure 2.2 shows a compilation of the weight versus the cruising speed for various aircraft (top right), birds (middle), insects, and butterflies (bottom left). This figure is reproduced from the wonderful book of Tennekes [106] about the science of flight. The points in this figure cluster around a straight line. The weight and the cruising speed are shown on a double logarithmic scale; hence the straight line implies a power law relation of the form $W \sim v^n$.

Problem e Measure the slope of the line in Figure 2.2 and show that this slope is close to the value $n = 6$ predicted by expression (2.31).

Note that the lift in Figure 2.2 ranges over 11 orders of magnitude. Despite this extreme range in parameter values, the scaling law (2.31) holds remarkably well. The individual points show departures from the scaling law. The reason is that the density ρ_f and the lift coefficient C_L vary among different flying objects; the shape of the wing of a Boeing 747 is different from the shape of the wing of a butterfly.

This example shows that dimensional arguments can be useful in explaining the relationship between different physical parameters. Such relationships are also of importance in the design of *scale experiments*. An example of a scale experiment is a model of an aircraft in a wind tunnel. All physical parameters need to be scaled appropriately with the size of the model aircraft so that the physics is unaltered by the scaling. This is the case when the dimensionless numbers determined with the Buckingham pi theorem are the same for the scaled model as for the real aircraft. In this way the Buckingham pi theorem provides a systematic procedure for the design of scale experiments as well [23].

2.6 Dependence of pipe flow on the radius of the pipe

The flow of a viscous fluid through a porous medium is important for understanding and managing aquifers and hydrocarbon reservoirs. Here we use dimensional analysis to study the dependence of flow of a viscous fluid through a cylindrical pipe as shown in Figure 2.3. The flow is driven by a pressure gradient $\partial p/\partial x$ along

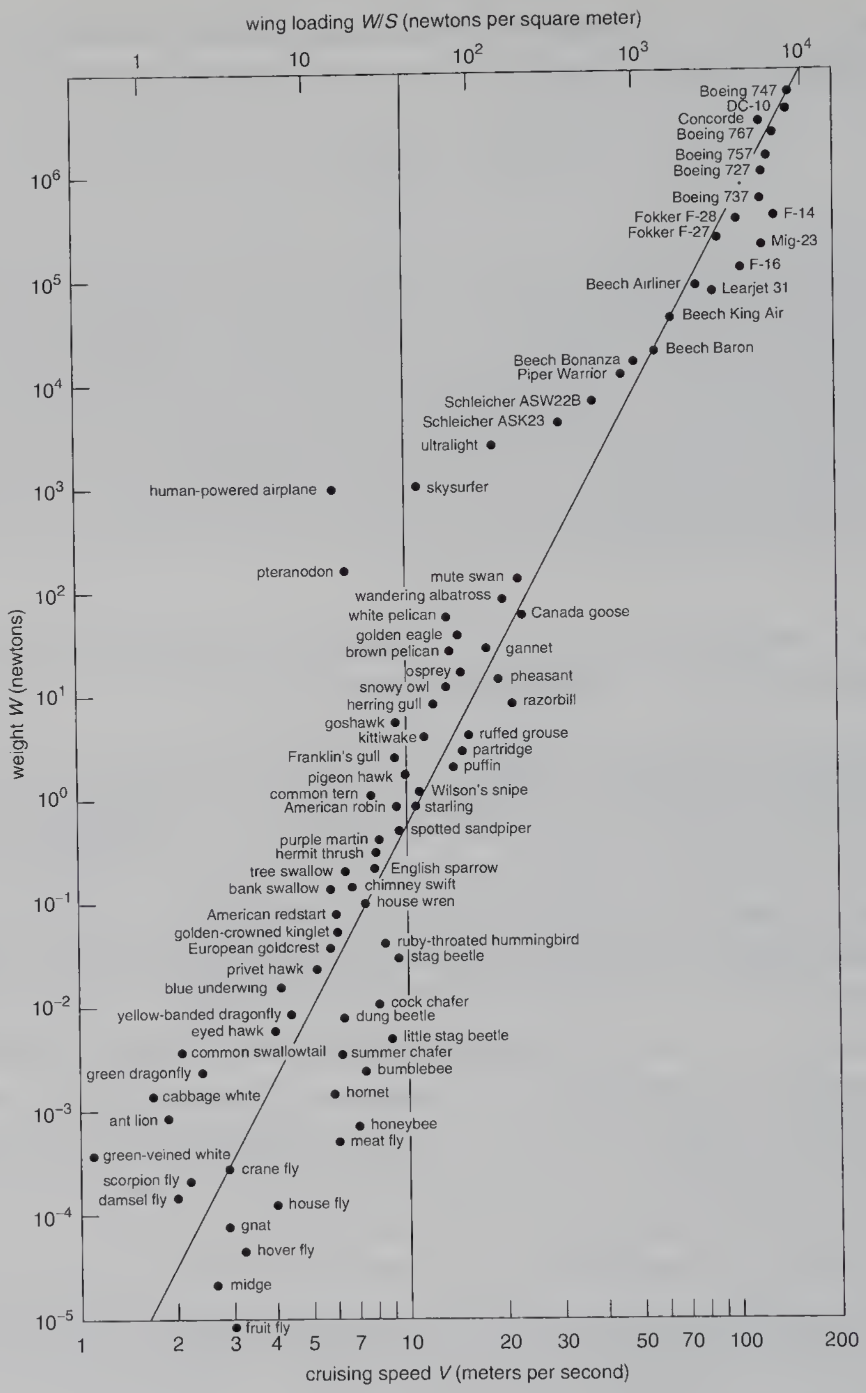

Fig. 2.2 The weight of many flying objects (vertical axis) against their cruising speed (horizontal axis) on a log–log plot. This figure is reproduced from reference [106] with permission from MIT Press.

the center axis of the cylinder. We assume that the fluid has a viscosity μ, and we want to find the relation between the strength of the flow along the pipe per unit time and the radius R. As a measure of the flow rate we use the volume of the flow per unit time, and designate this quantity with the symbol Φ.

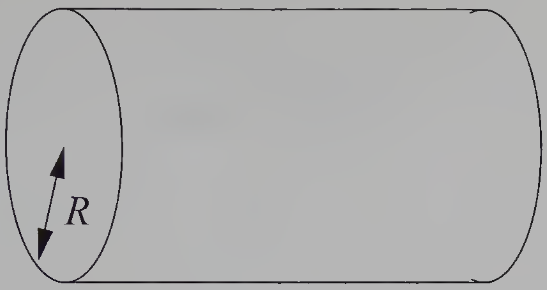

Fig. 2.3 The geometry of a pipe through which fluid flows.

Problem a The physical quantities that are of relevance to this problem are the pressure gradient $\partial p/\partial x$, the viscosity μ, the radius R, and the flow rate Φ. Write down the physical dimensions of each of these properties. In order to find the dimension of the viscosity you can use the relation $\tau = \mu \partial v/\partial z$, where τ is the shear stress (with dimension pressure), v the velocity, and z distance.

Problem b Use the Buckingham pi theorem to show that the flow rate is given by

$$\Phi = constant \, \frac{\partial p/\partial x}{\mu} R^4. \tag{2.32}$$

Problem c This expression states that the flow rate is proportional to the pressure gradient, which reflects the fact that a stronger pressure gradient generates a stronger driving force for the flow, and hence a stronger flow. Give a similar physical explanation for the dependence of the flow rate on the viscosity and the radius. At first you might think that the flow rate is proportional to the surface area πR^2 of the pipe. Try to give a physical explanation for the R^4-dependence of the flow rate on the radius.

The result (2.32) can also be obtained by solving the Navier–Stokes equation (11.55) for the appropriate boundary condition, and by integrating the flow velocity over the pipe to give the flow rate Φ. This treatment is more cumbersome than the analysis of this section, but it does provide the proportionality constant in expression (2.32).

3

Power series

3.1 Taylor series

In many applications in mathematical physics it is useful to write the quantity of interest as a sum of a number of terms. To fix our mind, let us consider the motion of a particle that moves along a line as time progresses. The motion is completely described by giving the position $x(t)$ of the particle as a function of time. Consider the four different types of motion that are shown in Figure 3.1.

The simplest motion is a particle that does not move, as shown in panel (a). In this case the position of the particle is constant:

$$x(t) = x_0. \tag{3.1}$$

The value of the parameter x_0 follows by setting $t = 0$ in this expression; this immediately gives

$$x_0 = x\,(0)\,. \tag{3.2}$$

In panel (b) the situation for a particle that moves with a constant velocity is shown, thus the position is a linear function of time:

$$x(t) = x_0 + v_0 t\,. \tag{3.3}$$

Again, setting $t = 0$ gives the parameter x_0, which is given again by (3.2). The value of the parameter v_0 follows by differentiating (3.3) with respect to time and by setting $t = 0$.

Problem a Do this and show that

$$v_0 = \frac{dx}{dt}(t = 0). \tag{3.4}$$

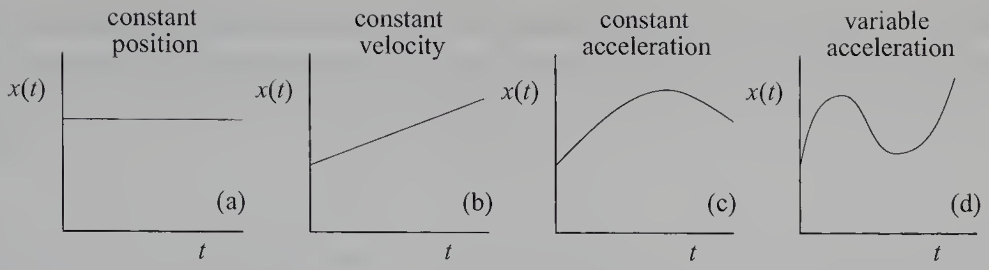

Fig. 3.1 Four different kinds of motion of a particle along a line as a function of time.

This expression reflects that the velocity v_0 is given by the time-derivative of the position. Next, consider a particle moving with a constant acceleration a_0 as shown in panel (c). As you probably know from classical mechanics the motion in that case is a quadratic function of time:

$$x(t) = x_0 + v_0 t + \frac{1}{2} a_0 t^2. \tag{3.5}$$

Problem b Evaluate this expression at $t = 0$ to show that x_0 is given by (3.2). Differentiate (3.5) once with respect to time and evaluate the result at $t = 0$ to show that v_0 is again given by (3.4). Differentiate (3.5) twice with respect to time, set $t = 0$ to show that a_0 is given by

$$a_0 = \frac{d^2 x}{dt^2}(t = 0). \tag{3.6}$$

This result reflects the fact that the acceleration is the second derivative of the position with respect to time.

Let us now consider the motion shown in panel (d) where the acceleration changes with time. In that case the displacement as a function of time is not a linear function of time (as in (3.3) for the case of a constant velocity) nor is it a quadratic function of time (as in (3.5) for the case of a constant acceleration). Instead, the displacement is in general a function of all possible powers in t:

$$x(t) = c_0 + c_1 t + c_2 t^2 + \cdots = \sum_{n=0}^{\infty} c_n t^n. \tag{3.7}$$

This series, in which a function is expressed as a sum of terms with increasing powers of the independent variable, is called a *Taylor series*. At this point we do not know what the constants c_n are. These coefficients can be found in exactly the same way as in problem b in which you determined the coefficients a_0 and v_0 in the expansion (3.5).

Problem c Determine the coefficient c_m by differentiating expression (3.7) m times with respect to t and by evaluating the result at $t = 0$ to show that

$$c_m = \frac{1}{m!}\frac{d^m x}{dt^m}(x = 0). \tag{3.8}$$

Of course there is no reason why the Taylor series can only be used to describe the displacement $x(t)$ as a function of time t. In the literature, the Taylor series is frequently used to describe a function $f(x)$ that depends on x. Of course it is immaterial what we call a function. By making the replacements $x \rightarrow f$ and $t \rightarrow x$ expressions (3.7) and (3.8) can also be written as:

$$f(x) = \sum_{n=0}^{\infty} c_n x^n, \tag{3.9}$$

with

$$c_n = \frac{1}{n!}\frac{d^n f}{dx^n}(x = 0). \tag{3.10}$$

You may also find this result in the literature written as

$$f(x) = \sum_{n=0}^{\infty} \frac{x^n}{n!}\frac{d^n f}{dx^n}(x = 0) = f(0) + x\frac{df}{dx}(x = 0) + \frac{x^2}{2}\frac{d^2 f}{dx^2}(x = 0) + \cdots. \tag{3.11}$$

Problem d By evaluating the derivatives of $f(x)$ at $x = 0$ show that the Taylor series of the following functions are given by:

$$\sin(x) = x - \frac{1}{3!}x^3 + \frac{1}{5!}x^5 - \cdots; \tag{3.12}$$

$$\cos(x) = 1 - \frac{1}{2}x^2 + \frac{1}{4!}x^4 - \cdots; \tag{3.13}$$

$$e^x = 1 + x + \frac{1}{2!}x^2 + \frac{1}{3!}x^3 + \cdots = \sum_{n=0}^{\infty} \frac{1}{n!}x^n; \tag{3.14}$$

$$\frac{1}{1-x} = 1 + x + x^2 + \cdots = \sum_{n=0}^{\infty} x^n; \tag{3.15}$$

$$(1-x)^\alpha = 1 - \alpha x + \frac{1}{2!}\alpha(\alpha - 1)x^2 - \frac{1}{3!}\alpha(\alpha - 1)(\alpha - 2)x^3 + \cdots. \tag{3.16}$$

Up to this point we made the Taylor expansion around the point $x = 0$. However, one can make a Taylor expansion of $f(x + h)$ around any arbitrary point x.

The associated Taylor series can be obtained by replacing the distance x that we move from the expansion point by a distance h and by replacing the expansion point 0 by x. Making the replacements $x \to h$ and $0 \to x$ expansion (3.11) is given by

$$f(x+h) = \sum_{n=0}^{\infty} \frac{h^n}{n!} \frac{d^n f}{dx^n}(x). \tag{3.17}$$

Problem e Truncate this series after the second term and show that this leads to the following approximations:

$$f(x+h) - f(x) \approx h \frac{df}{dx}(x), \tag{3.18}$$

$$\frac{df}{dx} \approx \frac{f(x+h) - f(x)}{h}. \tag{3.19}$$

These expressions may appear to be equivalent in a trivial way. However, we will make extensive use of them in different ways. Equation (3.18) makes it possible to estimate the change in a function when the independent variable is changed slightly, whereas (3.19) is useful for estimating the derivative of a function given its values at neighboring points. The issue of estimating the derivative of a function is treated in much more detail in Section 12.2. Figure 12.2 makes it possible to also derive the estimate (3.19) geometrically by using that the derivative of a function is just the slope of that function.

The Taylor series can also be used for functions of more than one variable. As an example consider a function $f(x, y)$ that depends on the variables x and y. The generalization of the Taylor series (3.9) to functions of two variables is given by

$$f(x, y) = \sum_{n,m=0}^{\infty} c_{nm} x^n y^m. \tag{3.20}$$

At this point the coefficients c_{nm} are not yet known. They follow in the same way as the coefficients of the Taylor series of a function that depends on a single variable by taking the partial derivatives of the Taylor series and evaluating the result at the point where the expansion is made.

Problem f Take all the partial derivatives of (3.20) with respect to x and y up to second order, including the mixed derivative $\partial^2 f / \partial x \partial y$, and evaluate the result at the expansion point $x = y = 0$ to show that up to second order the

Taylor expansion (3.20) is given by

$$
\begin{aligned}
f(x,y) = f(0,0) &+ \frac{\partial f}{\partial x}(0,0)\,x + \frac{\partial f}{\partial y}(0,0)\,y \\
&+ \frac{1}{2}\frac{\partial^2 f}{\partial x^2}(0,0)\,x^2 + \frac{\partial^2 f}{\partial x \partial y}(0,0)\,xy \\
&+ \frac{1}{2}\frac{\partial^2 f}{\partial y^2}(0,0)\,y^2 + \cdots .
\end{aligned}
\tag{3.21}
$$

Problem g This is the Taylor expansion of $f(x, y)$ around the point $x = y = 0$. Make suitable substitutions in this result to show that the Taylor expansion around an arbitrary point (x, y) is given by

$$
\begin{aligned}
f(x+h_x, y+h_y) = f(x,y) &+ \frac{\partial f}{\partial x}(x,y)\,h_x + \frac{\partial f}{\partial y}(x,y)\,h_y \\
&+ \frac{1}{2}\frac{\partial^2 f}{\partial x^2}(x,y)\,h_x^2 + \frac{\partial^2 f}{\partial x \partial y}(x,y)\,h_x h_y \\
&+ \frac{1}{2}\frac{\partial^2 f}{\partial y^2}(x,y)\,h_y^2 + \cdots .
\end{aligned}
\tag{3.22}
$$

Let us return to the Taylor series (3.9) with the coefficients c_m given by (3.10). This series hides an intriguing result. Equations (3.9) and (3.10) suggest that a function $f(x)$ is specified for all values of its argument x when all the derivatives are known at a single point $x = 0$. This means that the global behavior of a function is completely contained in the properties of the function at a single point. In fact, this is not always true.

First, the series (3.9) is an infinite series, and the sum of infinitely many terms does not necessarily lead to a finite answer. As an example look at the series (3.15). A series can only converge when the terms go to zero as $n \to \infty$, because otherwise every additional term changes the sum. The terms in the series (3.15) are given by x^n; these terms only go to zero as $n \to \infty$ when $|x| < 1$. In general, the Taylor series (3.9) only *converges* when x is smaller than a certain critical value called the *radius of convergence*. Details on the criteria for the convergence of series can be found in for example Boas [19] or Butkov [24].

The second reason why the derivatives at one point do not necessarily constrain the function everywhere is that a function may change its character over the range of parameter values that is of interest. As an example let us return to a moving particle and consider a particle at position $x(t)$ that is at rest until a certain time t_0

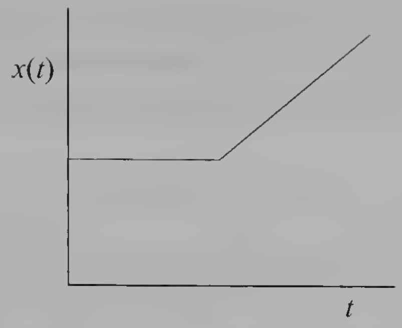

Fig. 3.2 The motion of a particle that suddenly changes character at time t_0.

and that then starts moving with a uniform velocity $v \neq 0$:

$$x(t) = \begin{cases} x_0 & \text{for } t \leq t_0 \\ x_0 + v(t - t_0) & \text{for } t > t_0 \end{cases}. \tag{3.23}$$

The motion of the particle is sketched in Figure 3.2. A straightforward application of (3.8) shows that all the coefficients c_n of this function vanish except c_0 which is given by x_0. The Taylor series (3.7) is therefore given by $x(t) = x_0$ which clearly differs from (3.23). The reason for this is that the function (3.23) changes its character at $t = t_0$ in such a way that nothing in the behavior for times $t < t_0$ predicts the sudden change in the motion at time $t = t_0$. Mathematically things go wrong because the first and higher derivatives of the function are not defined at time $t = t_0$.

Problem h What is the first derivative of $x(t)$ just to the left and just to the right of $t = t_0$? What is the second derivative at that point?

The function (3.23) is said to be not analytic at the point $t = t_0$. The issue of analytic functions is treated in more detail in Sections 16.1 and 17.1.

Problem i Try to compute the Taylor series of the function $x(t) = 1/t$ using (3.7) and (3.8). Draw this function and explain why the Taylor series cannot be used for this function.

Problem j Do the same for the function $x(t) = \sqrt{t}$.

The examples in the last two problems show that when a function is not analytic at a certain point, the coefficients of the Taylor series are not defined. This signals that such a function cannot be represented by a Taylor series around that point.

Frequently the result of a calculation can be obtained by summing a series. In Section 3.3 this is used to study the behavior of a bouncing ball. The bounces are "natural" units for analyzing the problem at hand. In Section 3.4 the reverse is done when studying the total reflection of a stack of reflective layers. In this case a series expansion actually gives physical insight into a complex expression.

3.2 Growth of the Earth by cosmic dust

In this section we use the growth of the Earth by the accretion of cosmic dust as an example to illustrate the usefulness of the (first order) Taylor series. The Earth is continuously bombarded from space by meteorites. Some of these meteorites can be large and lead to massive impact craters. As an example the gravity anomaly over the Chicxulub impact crater in Mexico is shown in Figure 21.1. The diameter of this impact crater is about 100 km. However, the bulk of the cosmic dust that falls from space onto the Earth is in the form of many small particles. The total mass of all the cosmic dust that falls on the Earth is estimated by Love and Brownlee [65] to be approximately 5×10^7 kg/a. (The unit "a" stands for annum (or year); this means that the unit used here is kilograms per year.) This estimate, however, is not very accurate and we can probably only trust the first decimal of this number. This means that in subsequent calculations it is pointless to aim for an accuracy of more than one significant figure.

Since the cosmic dust increases the mass of the Earth, the size of the Earth will increase. In this section we determine the growth of the Earth's radius per year due to the bombardment of our planet by cosmic dust.

Problem a Assuming that a density of meteorites is given by $\rho = 2.5 \times 10^3$ kg/m^3 [65] show that the annual growth of the volume of the Earth is given by

$$\delta V = 2 \times 10^4 \text{ m}^3. \tag{3.24}$$

Also show that this corresponds to a block of $(27 \times 27 \times 27)$ m^3.

We assume that the Earth is a perfect sphere so that the volume and the radius r of the Earth are related by the relation

$$V = \frac{4\pi}{3} r^3. \tag{3.25}$$

From this relation we can deduce that the annual change δr of the radius of the Earth can be computed from the expression

$$\delta r = \left[\frac{3(V + \delta V)}{4\pi}\right]^{1/3} - \left(\frac{3V}{4\pi}\right)^{1/3}. \tag{3.26}$$

Problem b Assume the Earth's radius is given by $r = 7000$ km. Insert this number and the value of δV from expression (3.24) into (3.26) and use a calculator to compute the increase, δr, of the radius of the Earth.

You have probably found that the annual increase in the Earth's radius is equal to zero. This cannot be true because we know that δV is not equal to zero. The reason that your calculator has given you a wrong answer is that in (3.26) we are subtracting two very large numbers that differ by a small amount. The volume V of the Earth is of the order 10^{21} m^3 while according to (3.24) the annual increase of the volume is of the order 10^4 m^3. Most calculators carry out all the calculations with a relatively small number of digits; most use between 6 and 10 decimals. When you subtract two numbers that are very large and that have a very small difference, this difference will be truncated after say 6 or 10 decimals. In our problem, the first ten decimals of both terms in (3.26) are identical, hence your calculator tells you that the radius of the Earth is not growing because of the accretion of cosmic dust.

In general, subtracting two large numbers that have a difference that is much smaller leads to numerical inaccuracies. Clearly a trick is needed to obtain the desired growth of the Earth's radius. The cause of this problem is that the annual change in the volume is so small. We can turn this problem to our advantage by using that the Taylor series that we introduced in the previous section is extremely accurate when the independent variable is changed by a very small amount. Here we will compute the increase of the Earth's radius using expression (3.18).

Problem c Show that this expression can also be written as

$$\delta f \approx \frac{\partial f}{\partial x}\delta x, \tag{3.27}$$

where δf is the change in the function $f(x)$ due to a change δx in the independent variable x.

Problem d Apply this result to the function $r(V) = (3V/4\pi)^{1/3}$ that gives the radius as a function of the volume to derive that

$$\delta r = \frac{1}{3} r \frac{\delta V}{V}. \tag{3.28}$$

Problem e Use this result to compute the annual increase of the radius of the Earth due to the accretion of cosmic dust and show that the result is of the order of 1 ångström per year (1 ångström is 10^{-10} m).

Problem f Can you think of an object that is the size of 1 ångström?

Problem g How much has the Earth's radius increased over the age of the Earth? In this calculation you may assume that the age of the Earth is 4.5 billion years.

The upshot of this calculation is that the growth of the Earth due to the present-day accretion of cosmic dust is negligible. However, the technique of using the first order Taylor series to determine the small change in a quantity is extremely powerful. In fact, you have encountered in this section an example that demonstrates that an approximation can provide a more meaningful answer than a calculation carried out using a calculator or computer.

3.3 Bouncing ball

In this section we study a rubber ball that bounces on a flat surface and slowly comes to rest as sketched in Figure 3.3. You will know from experience that the ball bounces more and more rapidly with time. The question we address here is whether the ball can actually bounce infinitely many times in a finite amount of time. This problem is not an easy one. In general with large difficult problems it is a useful strategy to divide the large and difficult problem that you cannot solve into smaller and simpler problems that you can solve. By assembling these smaller sub-problems one can then often solve the large problem. This is exactly what we will do here. First we will find how much time it takes for the ball to bounce once given its velocity. Given a prescription of the energy loss in one bounce we will determine a relation between the velocity of subsequent bounces. From these ingredients we can determine the relation between the times needed for subsequent bounces. By summing this series over an infinite number of bounces we can determine the total time that the ball has bounced. *Keep this general strategy in mind when solving complex problems. Almost all of us are better at solving a number of small problems rather than a single large problem!*

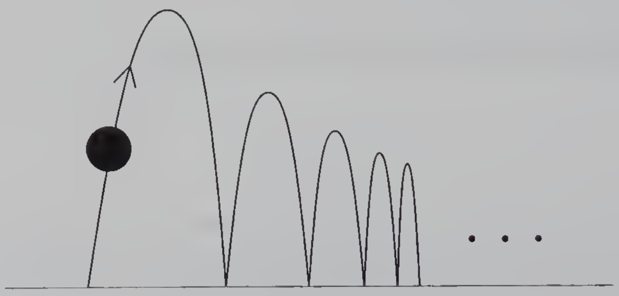

Fig. 3.3 The motion of a bouncing ball that loses energy with every bounce. To visualize the motion of the ball better, the ball is a given a constant horizontal velocity that is conserved during the bouncing.

Problem a A ball moves upward from the level $z = 0$ with velocity v and is subject to a constant gravitational acceleration g. Determine the height the ball reaches and the time it takes for the ball to return to its starting point.

At this point we have determined the relevant properties for a single bounce. During each bounce the ball loses energy due to the fact that the ball is deformed inelastically during the bounce. We assume that during each bounce the ball loses a fraction γ of its energy.

Problem b Let the velocity at the beginning of the nth bounce be v_n. Show that with the assumed rule for energy loss this velocity is related to the velocity v_{n-1} of the previous bounce by

$$v_n = \sqrt{1-\gamma}\ v_{n-1}. \tag{3.29}$$

Hint: when the ball bounces upward from $z = 0$ all its energy is kinetic energy $\frac{1}{2}mv^2$.

In problem a you determined the time it took the ball to bounce once, given the initial velocity, while expression (3.29) gives a recursive relation for the velocity between subsequent bounces. In problem a you also computed the time that it takes to carry out a single bounce. By assembling these results we can find a relation between the time t_n for the nth bounce and the time t_{n-1} for the previous bounce.

Problem c Determine this relation. In addition, let us assume that the ball is thrown up the first time from $z = 0$ to reach a height $z = H$. Compute the time t_0 needed for the ball to make the first bounce and combine these results to show that

$$t_n = \sqrt{\frac{8H}{g}}(1-\gamma)^{n/2}, \tag{3.30}$$

where g is the acceleration of gravity.

We can use this expression to determine the total time T_N it takes to carry out N bounces. This time is given by $T_N = \sum_{n=0}^{N} t_n$. By setting N equal to infinity we can compute the time T_∞ it takes to bounce infinitely often.

Problem d Determine this time by carrying out the summation and show that it is given by:

$$T_\infty = \sqrt{\frac{8H}{g}} \frac{1}{1-\sqrt{1-\gamma}}. \tag{3.31}$$

Hint: write $(1-\gamma)^{n/2}$ as $\left(\sqrt{1-\gamma}\right)^n$ and treat $\sqrt{1-\gamma}$ as the parameter x in the appropriate Taylor series of Section 3.1.

This result shows that the time it takes to bounce infinitely often is indeed finite. For the special case that the ball loses no energy, $\gamma = 0$ and T_∞ is infinite. This reflects that a ball that loses no energy will bounce forever.

Expression (3.31) looks messy. It often happens in mathematical physics that the final expression resulting from a calculation is so complex that it is difficult to understand it. However, often we know that certain terms in an expression can be assumed to be very small (or very large). This may allow us to obtain an approximate expression that is of a simpler form. In this way we trade accuracy for simplicity and understanding. In practice, this often turns out to be a good deal! In our example of the bouncing ball we assume that the energy loss at each bounce is small, that is that γ is small.

Problem e Show that in this case $T_\infty \approx \sqrt{(8H/g)}\, 2/\gamma$ by using the leading terms of the appropriate Taylor series of Section 3.1.

This result is actually quite useful. It tells us *how* the total bounce time approaches infinity when the energy loss γ goes to zero.

In this example we have solved the problem in little steps. In general we take larger steps in the problems in this book, and you will have to discover how to divide a large step into smaller steps. The next problem is a "large" problem; solve it by dividing it into smaller problems. First formulate the smaller problems as ingredients for the large problem before you actually start working on the smaller problems.

> *Make it a habit whenever you are solving problems to first formulate a strategy for how you are going to attack a problem before you actually start working on the sub-problems. Make a list if this helps you and do not be deterred if you cannot solve a particular sub-problem. Perhaps you can solve the other sub-problems and somebody else can help you with the one you cannot solve.*

Keeping this in mind, solve the following "large" problem:

Problem f Let the total distance traveled by the ball in the vertical direction during infinitely many bounces be denoted by S. Show that $S = 2H/\gamma$.

3.4 Reflection and transmission by a stack of layers

In 1917 Lord Rayleigh [86] addressed the question of why some birds and insects have beautiful iridescent colors. He expiained this by studying the reflective properties of a stack of thin reflective layers. This problem is also of interest in geophysics; in exploration seismology one is also interested in the reflection and transmission properties of stacks of reflective layers in the Earth. Lord Rayleigh solved this problem in the following way.

Suppose we have one stack of layers on the left with reflection coefficient R_L and transmission coefficient T_L and another stack of layers on the right with reflection coefficient R_R and transmission coefficient T_R. If we add these two stacks together to obtain a larger stack of layers, what are the reflection coefficient R and transmission coefficient T of the total stack of layers? See Figure 3.4 for the scheme of this problem. The reflection coefficient is defined as the ratio of the strengths of the reflected and the incident waves, similarly the transmission coefficient is defined as the ratio of the strengths of the transmitted wave and the incident wave. To highlight the essential arguments we simplify the analysis and ignore that the reflection coefficient for waves incident from the left and the right are in general not the same. However, this simplification does not change the essence of the coming arguments.

Before we start solving the problem, let us speculate what the transmission coefficient of the combined stack is. It may seem natural to assume that the transmission

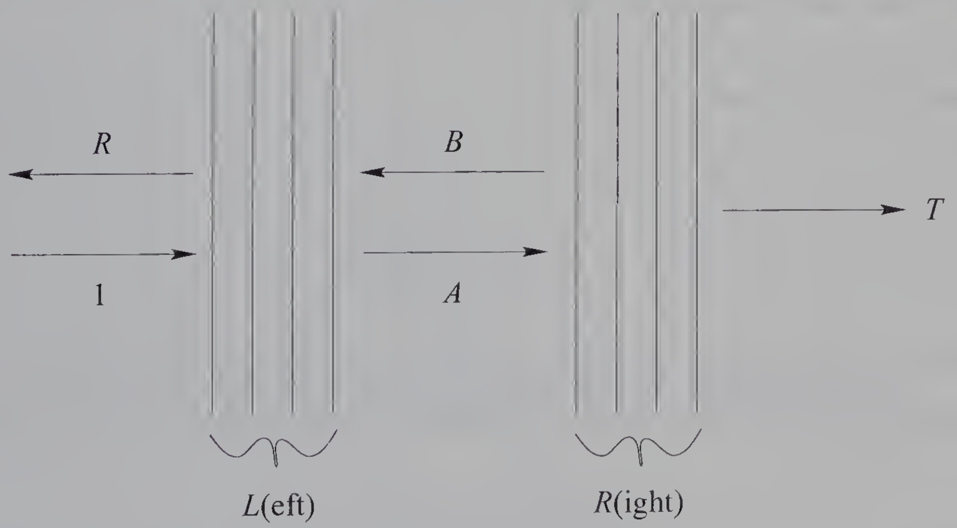

Fig. 3.4 Geometry of the problem where stacks of n and m reflective layers are combined. The notation of the strength of left- and right-going waves is indicated.

coefficient of the combined stack is the product of the transmission coefficient of the individual stacks:

$$T \overset{?}{=} T_L T_R. \tag{3.32}$$

However, this result is wrong and we will discover why this is so. Consider Figure 3.4 again. The unknown quantities are R, T, and the coefficients A and B for the right-going and left-going waves between the stacks. An incident wave with strength 1 impinges on the stack from the left. Let us first determine the coefficient A of the right-going waves between the stacks. The right-going wave between the stacks contains two contributions: the wave transmitted from the left (this contribution has a strength $1 \times T_L$) and the wave reflected towards the right due the incident left-going wave with strength B (this contribution has a strength $B \times R_L$). This implies that:

$$A = T_L + BR_L. \tag{3.33}$$

Problem a Using similar arguments show that:

$$B = AR_R, \tag{3.34}$$

$$T = AT_R, \tag{3.35}$$

$$R = R_L + BT_L. \tag{3.36}$$

This is all we need to solve our problem. The system of equations (3.33)–(3.36) consists of four linear equations with four unknowns A, B, R, and T. We could solve this system of equations by brute force, but some thought will make life easier for us. Note that the last two equations immediately give T and R once A and B are known. The first two equations give A and B.

Problem b Show that

$$A = \frac{T_L}{(1 - R_L R_R)}, \tag{3.37}$$

$$B = \frac{T_L R_R}{(1 - R_L R_R)}. \tag{3.38}$$

This is a puzzling result, the right-going wave A between the layers not only contains the transmission coefficient of the left layer T_L but also an additional term $1/(1 - R_L R_R)$.

Problem c Make a series expansion of $1/(1 - R_L R_R)$ in the quantity $R_L R_R$ and show that this term accounts for the waves that bounce back and forth between the two stacks. Hint: use that R_L is the reflection coefficient for a wave that reflects from the left stack and R_R is the reflection coefficient for one that reflects from the right stack so that $R_L R_R$ is the total reflection coefficient for a wave that bounces once between the left and the right stacks.

This implies that the term $1/(1 - R_L R_R)$ accounts for the waves that bounce back and forth between the two stacks of layers. It is for this reason that we call this a *reverberation* term. It plays an important role in computing the response of layered media.

Problem d Show that the reflection and transmission coefficients of the combined stack of layers are given by:

$$R = R_L + \frac{T_L^2 R_R}{(1 - R_L R_R)}, \tag{3.39}$$

$$T = \frac{T_L T_R}{(1 - R_L R_R)}. \tag{3.40}$$

At the beginning of this section we conjectured that the transmission coefficient of the combined stacks is the product of the transmission coefficient of the separate stacks, see expression (3.32).

Problem e Is this conjecture correct? Under which conditions is it approximately correct?

Equations (3.39) and (3.40) are useful for computing the reflection and transmission coefficients of a large stack of layers. The reason for this is that it is extremely simple to determine the reflection and transmission coefficients of a very thin layer using the Born approximation. (The Born approximation is treated in Section 23.2.) Let the reflection and transmission coefficients of a *single* thin layer n be denoted by r_n and t_n respectively and let the reflection and transmission coefficients of a *stack* of n layers be denoted by R_n and T_n respectively. Suppose that the left stack consists of n layers and that we want to add an $(n + 1)$th layer to the stack. In that case the right stack consists of a single $(n + 1)$th layer so that $R_R = r_{n+1}$ and $T_R = t_{n+1}$ and the reflection and transmission coefficients of the left stack are given by $R_L = R_n$, $T_L = T_n$. Using this in expressions (3.39) and (3.40) yields

$$R_{n+1} = R_n + \frac{T_n^2 r_{n+1}}{(1 - R_n r_{n+1})}, \tag{3.41}$$

$$T_{n+1} = \frac{T_n t_{n+1}}{(1 - R_n r_{n+1})}. \tag{3.42}$$

This means that given the known response of a stack of n layers, one can easily compute the effect of adding the $(n+1)$th layer to this stack. In this way one can recursively build up the response of the complex reflector out of the known response of very thin reflectors. Computers are pretty stupid, but they are ideally suited for applying the rules (3.41) and (3.42) a large number of times. Of course this process has to be begun with a medium in which no layers are present.

Problem f What are the reflection coefficient R_0 and the transmission coefficient T_0 when there are as yet no reflective layers present? Describe how one can compute the response of a thick stack of layers once we know the response of a very thin layer.

In developing this theory, Lord Rayleigh prepared the foundations for a theory that later became known as *invariant embedding* which turns out to be extremely useful for a number of scattering and diffusion problems [13, 109].

The main conclusion of the treatment of this section is that the transmission of a combination of two stacks of layers is not the product of the transmission coefficients of the two separate stacks because the waves that repeatedly reflect between the two stacks leave an imprint on the transmission coefficient as well. Paradoxically, Berry and Klein [15] showed in their analysis of "transparent mirrors" that for the special case of a large stack of layers with random transmission coefficient the total transmission coefficient *is* the product of the transmission coefficients of the individual layers, despite the fact that multiple reflections play a crucial role in this process.

4

Spherical and cylindrical coordinates

Many problems in mathematical physics exhibit a spherical or cylindrical symmetry. For example, the gravity field of the Earth is to first order spherically symmetric. Waves excited by a stone thrown into water are usually cylindrically symmetric. Although there is no reason why problems with such a symmetry cannot be analyzed using Cartesian coordinates (i.e. (x, y, z)-coordinates), it is usually not very convenient to use such a coordinate system. The reason for this is that the theory is usually much simpler when one selects a coordinate system with symmetry properties that are the same as the symmetry properties of the physical system that one wants to study. It is for this reason that spherical coordinates and cylindrical coordinates are introduced in this section. It takes a certain effort to become acquainted with these coordinate systems, but this effort is well spent because it makes solving a large class of problems much easier.

4.1 Introducing spherical coordinates

In Figure 4.1 a Cartesian coordinate system with its x-, y-, and z-axes is shown as well as the location of a point $\mathbf{r}$. This point can be described either by its x-, y-, and z-components or by the radius r and the angles θ and φ shown in Figure 4.1. In the latter case one uses spherical coordinates. Comparing the angles θ and φ with the geographical coordinates that define a point on the globe one sees that φ can be compared with *longitude* and θ can be compared with *colatitude*, which is defined as (*latitude* $-$ 90 degrees).

Problem a The city of Utrecht in the Netherlands is located at 52 degrees north and 5 degrees east. Show that the angles θ and φ (in radians) that correspond to this point on the sphere are given by $\theta = 0.663$ radians, and $\varphi = 0.087$ radians.

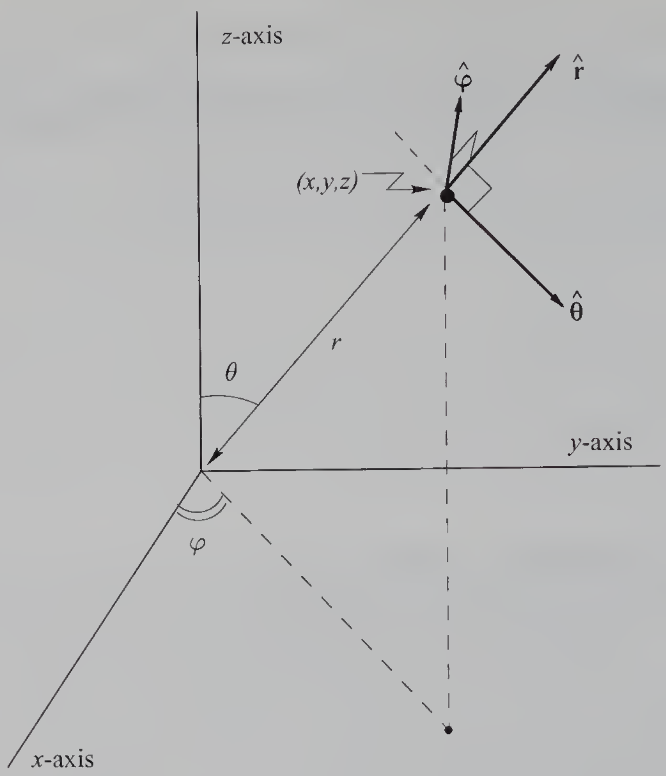

Fig. 4.1 Definition of the angles used in the spherical coordinates.

The angle φ runs from 0 to 2π, while θ has values between 0 and π. In terms of Cartesian coordinates the position vector can be written as:

$$\mathbf{r} = x\hat{\mathbf{x}} + y\hat{\mathbf{y}} + z\hat{\mathbf{z}}, \tag{4.1}$$

where the caret (ˆ) is used to denote a vector that is of unit length. An arbitrary vector can be expressed in a superposition of these basis vectors:

$$\mathbf{u} = u_x\hat{\mathbf{x}} + u_y\hat{\mathbf{y}} + u_z\hat{\mathbf{z}}. \tag{4.2}$$

We want also to express the same vector in basis vectors that are related to the spherical coordinate system. Before we can do so we must first establish the connection between the Cartesian coordinates (x, y, z) and the spherical coordinates (r, θ, φ).

Problem b Use Figure 4.1 to show that the Cartesian coordinates are given by:

$$\left.\begin{aligned} x &= r\sin\theta\cos\varphi, \\ y &= r\sin\theta\sin\varphi, \\ z &= r\cos\theta. \end{aligned}\right\} \tag{4.3}$$

Problem c Use these expressions to derive the following expression for the spherical coordinates in terms of the Cartesian coordinates:

$$\left.\begin{aligned} r &= \sqrt{x^2+y^2+z^2}, \\ \theta &= \arccos\left(z/\sqrt{x^2+y^2+z^2}\right), \\ \varphi &= \arctan\left(y/x\right). \end{aligned}\right\} \tag{4.4}$$

We have now obtained the relation between the Cartesian coordinates (x, y, z) and the spherical coordinates (r, θ, φ). Suppose we want to express the vector $\mathbf{u}$ of equation (4.2) in spherical coordinates:

$$\mathbf{u} = u_r\hat{\mathbf{r}} + u_\theta\hat{\boldsymbol{\theta}} + u_\varphi\hat{\boldsymbol{\varphi}}, \tag{4.5}$$

and we want to know the relation between the components (u_x, u_y, u_z) in Cartesian coordinates and the components $(u_r, u_\theta, u_\varphi)$ of the same vector expressed in spherical coordinates. In order to do this we first need to determine the unit vectors $\hat{\mathbf{r}}$, $\hat{\boldsymbol{\theta}}$, and $\hat{\boldsymbol{\varphi}}$. In Cartesian coordinates, the unit vector $\hat{\mathbf{x}}$ points along the x-axis. This is a different way of saying that it is a unit vector pointing in the direction of increasing values of x for constant values of y and z; in other words, $\hat{\mathbf{x}}$ can be written as: $\hat{\mathbf{x}} = \partial\mathbf{r}/\partial x$.

Problem d Verify this by showing that the differentiation $\hat{\mathbf{x}} = \partial\mathbf{r}/\partial x$ leads to the correct unit vector in the x-direction: $\hat{\mathbf{x}} = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}$.

Now consider the unit vector $\hat{\boldsymbol{\theta}}$. Using the same argument as for the unit vector $\hat{\mathbf{x}}$ we know that $\hat{\boldsymbol{\theta}}$ is directed towards increasing values of θ for constant values of r and φ. This means that $\hat{\boldsymbol{\theta}}$ can be written as $\hat{\boldsymbol{\theta}} = C\partial\mathbf{r}/\partial\boldsymbol{\theta}$. The constant C follows from the requirement that $\hat{\boldsymbol{\theta}}$ is of unit length.

Problem e Use this reasoning for all the unit vectors $\hat{\mathbf{r}}$, $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\varphi}}$ and expression (4.3) to show that:

$$\hat{\mathbf{r}} = \frac{\partial\mathbf{r}}{\partial r}, \quad \hat{\boldsymbol{\theta}} = \frac{1}{r}\frac{\partial\mathbf{r}}{\partial\theta}, \quad \hat{\boldsymbol{\varphi}} = \frac{1}{r\sin\theta}\frac{\partial\mathbf{r}}{\partial\varphi}, \tag{4.6}$$

and that this result can also be written as

$$\hat{\mathbf{r}} = \begin{pmatrix} \sin\theta\cos\varphi \\ \sin\theta\sin\varphi \\ \cos\theta \end{pmatrix}, \quad \hat{\boldsymbol{\theta}} = \begin{pmatrix} \cos\theta\cos\varphi \\ \cos\theta\sin\varphi \\ -\sin\theta \end{pmatrix}, \quad \hat{\boldsymbol{\varphi}} = \begin{pmatrix} -\sin\varphi \\ \cos\varphi \\ 0 \end{pmatrix}. \tag{4.7}$$

These equations give the unit vectors $\hat{\mathbf{r}}$, $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\varphi}}$ in Cartesian coordinates.

On the right-hand side of (4.6) the derivatives of the position vector are divided by 1, r, and $r\sin\theta$ respectively. These factors are usually shown in the following notation:

$$h_r = 1, \quad h_\theta = r, \quad h_\varphi = r\sin\theta. \tag{4.8}$$

These scale factors play an important role in the general theory of curvilinear coordinate systems, see Butkov [24] for details. The material presented in the remainder of this chapter as well as the derivation of vector calculus in spherical coordinates can be based on the scale factors given in (4.8). However, this approach will not be taken here.

Problem f Verify explicitly that the vectors $\hat{\mathbf{r}}$, $\hat{\boldsymbol{\theta}}$, and $\hat{\boldsymbol{\varphi}}$ defined in this way form an orthonormal basis, that is they are of unit length and perpendicular to each other:

$$(\hat{\mathbf{r}}\cdot\hat{\mathbf{r}}) = \left(\hat{\boldsymbol{\theta}}\cdot\hat{\boldsymbol{\theta}}\right) = (\hat{\boldsymbol{\varphi}}\cdot\hat{\boldsymbol{\varphi}}) = 1, \tag{4.9}$$

$$\left(\hat{\mathbf{r}}\cdot\hat{\boldsymbol{\theta}}\right) = (\hat{\mathbf{r}}\cdot\hat{\boldsymbol{\varphi}}) = \left(\hat{\boldsymbol{\theta}}\cdot\hat{\boldsymbol{\varphi}}\right) = 0. \tag{4.10}$$

The dot denotes the inner product of two vectors.

Problem g Using expressions (4.7) for the unit vectors $\hat{\mathbf{r}}$, $\hat{\boldsymbol{\theta}}$, and $\hat{\boldsymbol{\varphi}}$ show by calculating the cross-products explicitly that

$$\hat{\mathbf{r}}\times\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\varphi}}, \quad \hat{\boldsymbol{\theta}}\times\hat{\boldsymbol{\varphi}} = \hat{\mathbf{r}}, \quad \hat{\boldsymbol{\varphi}}\times\hat{\mathbf{r}} = \hat{\boldsymbol{\theta}}. \tag{4.11}$$

The Cartesian basis vectors $\hat{\mathbf{x}}$, $\hat{\mathbf{y}}$, and $\hat{\mathbf{z}}$ point in the same direction at every point in space. This is not true for the spherical basis vectors $\hat{\mathbf{r}}$, $\hat{\boldsymbol{\theta}}$, and $\hat{\boldsymbol{\varphi}}$; for different values of the angles θ and φ these vectors point in different directions. This implies that these unit vectors are functions of both θ and φ. For several applications it is necessary to know how the basis vectors change with θ and φ. This change is described by the derivative of the unit vectors with respect to the angles θ and φ.

Problem h Show by direct differentiation of expressions (4.7) that the derivatives of the unit vectors with respect to the angles θ and φ are given by:

$$\left.\begin{array}{ll} \partial\hat{\mathbf{r}}/\partial\theta = \hat{\boldsymbol{\theta}}, & \partial\hat{\mathbf{r}}/\partial\varphi = \sin\theta\ \hat{\boldsymbol{\varphi}}, \\ \partial\hat{\boldsymbol{\theta}}/\partial\theta = -\hat{\mathbf{r}}, & \partial\hat{\boldsymbol{\theta}}/\partial\varphi = \cos\theta\ \hat{\boldsymbol{\varphi}}, \\ \partial\hat{\boldsymbol{\varphi}}/\partial\theta = 0, & \partial\hat{\boldsymbol{\varphi}}/\partial\varphi = -\sin\theta\ \hat{\mathbf{r}} - \cos\theta\ \hat{\boldsymbol{\theta}}. \end{array}\right\} \tag{4.12}$$

4.2 Changing coordinate systems

Now that we have derived the properties of the unit vectors $\hat{\mathbf{r}}$, $\hat{\boldsymbol{\theta}}$, and $\hat{\boldsymbol{\varphi}}$, we are in the position to derive how the components $(u_r, u_\theta, u_\varphi)$ of the vector $\mathbf{u}$ defined in equation (4.5) are related to the usual Cartesian coordinates (u_x, u_y, u_z). This can most easily be achieved by writing expressions (4.7) in the following form:

$$\left.\begin{aligned}
\hat{\mathbf{r}} &= \sin\theta\cos\varphi\;\hat{\mathbf{x}} + \sin\theta\sin\varphi\;\hat{\mathbf{y}} + \cos\theta\;\hat{\mathbf{z}},\\
\hat{\boldsymbol{\theta}} &= \cos\theta\cos\varphi\;\hat{\mathbf{x}} + \cos\theta\sin\varphi\;\hat{\mathbf{y}} - \sin\theta\;\hat{\mathbf{z}},\\
\hat{\boldsymbol{\varphi}} &= -\sin\varphi\;\hat{\mathbf{x}} + \cos\varphi\;\hat{\mathbf{y}}.
\end{aligned}\right\} \tag{4.13}$$

Problem a Convince yourself that this expression can also be written in a symbolic form as

$$\begin{pmatrix} \hat{\mathbf{r}} \\ \hat{\boldsymbol{\theta}} \\ \hat{\boldsymbol{\varphi}} \end{pmatrix} = \mathbf{M} \begin{pmatrix} \hat{\mathbf{x}} \\ \hat{\mathbf{y}} \\ \hat{\mathbf{z}} \end{pmatrix}, \tag{4.14}$$

with the matrix $\mathbf{M}$ given by

$$\mathbf{M} = \begin{pmatrix} \sin\theta\cos\varphi & \sin\theta\sin\varphi & \cos\theta \\ \cos\theta\cos\varphi & \cos\theta\sin\varphi & -\sin\theta \\ -\sin\varphi & \cos\varphi & 0 \end{pmatrix}. \tag{4.15}$$

Of course expression (4.14) can only be considered to be a shorthand notation for equations (4.13) since the entries in (4.14) are vectors rather than single components.

The relation between the spherical components $(u_r, u_\theta, u_\varphi)$ and the Cartesian components (u_x, u_y, u_z) of the vector $\mathbf{u}$ can be obtained by inserting expressions (4.13) for the spherical coordinate unit vectors into the relation $\mathbf{u} = u_r\hat{\mathbf{r}} + u_\theta\hat{\boldsymbol{\theta}} + u_\varphi\hat{\boldsymbol{\varphi}}$.

Problem b Do this and collect together all terms multiplying the unit vectors $\hat{\mathbf{x}}$, $\hat{\mathbf{y}}$ and $\hat{\mathbf{z}}$ to show that expression (4.5) for the vector $\mathbf{u}$ is equivalent to:

$$\begin{aligned}
\mathbf{u} = {}& \left(u_r\sin\theta\cos\varphi + u_\theta\cos\theta\cos\varphi - u_\varphi\sin\varphi\right)\hat{\mathbf{x}}\\
&+ \left(u_r\sin\theta\sin\varphi + u_\theta\cos\theta\sin\varphi + u_\varphi\cos\varphi\right)\hat{\mathbf{y}}\\
&+ (u_r\cos\theta - u_\theta\sin\theta)\,\hat{\mathbf{z}}.
\end{aligned} \tag{4.16}$$

Problem c Show that this relation can also be written as:

$$\begin{pmatrix} u_x \\ u_y \\ u_z \end{pmatrix} = \mathbf{M}^T \begin{pmatrix} u_r \\ u_\theta \\ u_\varphi \end{pmatrix}, \tag{4.17}$$

where the matrix **M** is given by (4.15). In this expression, $\mathbf{M}^T$ is the transpose of the matrix **M**; that is it is the matrix obtained by interchanging rows and columns of the matrix $M_{ij}^T = M_{ji}$.

We have not yet reached with equation (4.17) our goal of expressing the spherical coordinate components $(u_r, u_\theta, u_\varphi)$ of the vector **u** in the Cartesian components (u_x, u_y, u_z). This is most easily achieved by multiplying (4.17) with the inverse matrix $(\mathbf{M}^T)^{-1}$, which gives:

$$\begin{pmatrix} u_r \\ u_\theta \\ u_\varphi \end{pmatrix} = \left(\mathbf{M}^T\right)^{-1} \begin{pmatrix} u_x \\ u_y \\ u_z \end{pmatrix} . \tag{4.18}$$

However, now we have only shifted the problem because we do not know the inverse $(\mathbf{M}^T)^{-1}$. We could of course painstakingly compute this inverse, but this would be a laborious process that we can avoid. It follows by inspection of (4.15) that all the columns of **M** are of unit length and that the columns are orthogonal. This implies that **M** is an orthogonal matrix. Orthogonal matrices have the useful property that the transpose of the matrix is identical to the inverse of the matrix: $\mathbf{M}^{-1} = \mathbf{M}^T$.

Problem d The property $\mathbf{M}^{-1} = \mathbf{M}^T$ can be verified explicitly by showing that $\mathbf{M}\mathbf{M}^T$ and $\mathbf{M}^T\mathbf{M}$ are equal to the identity matrix; do this!

Note that we have obtained the inverse of the matrix by making a guess and by verifying that this guess indeed solves our problem. This approach is often very useful in solving mathematical problems; there is nothing wrong with making a guess as long as you check afterwards that your guess is indeed a solution to your problem. Since we know that $\mathbf{M}^{-1} = \mathbf{M}^T$, it follows that $(\mathbf{M}^T)^{-1} = (\mathbf{M}^{-1})^{-1} = \mathbf{M}$.

Problem e Use these results to show that the spherical coordinate components of **u** are related to the Cartesian coordinates by the following transformation rule:

$$\begin{pmatrix} u_r \\ u_\theta \\ u_\varphi \end{pmatrix} = \begin{pmatrix} \sin\theta\cos\varphi & \sin\theta\sin\varphi & \cos\theta \\ \cos\theta\cos\varphi & \cos\theta\sin\varphi & -\sin\theta \\ -\sin\varphi & \cos\varphi & 0 \end{pmatrix} \begin{pmatrix} u_x \\ u_y \\ u_z \end{pmatrix} . \tag{4.19}$$

4.3 Acceleration in spherical coordinates

You may wonder whether we really need all these transformation rules between a Cartesian coordinate system and a system of spherical coordinates. The answer is yes! An important example can be found in meteorology where air moves along a spherical surface. The velocity **v** of the air can be expressed in spherical coordinates as:

$$\mathbf{v} = v_r \hat{\mathbf{r}} + v_\theta \hat{\boldsymbol{\theta}} + v_\varphi \hat{\boldsymbol{\varphi}}. \tag{4.20}$$

The motion of the air is governed by Newton's law, but when the velocity **v** and the force **F** are both expressed in spherical coordinates it would be wrong to express the θ-component of Newton's law as: $\rho dv_\theta/dt = F_\theta$. The reason is that the basis vectors of the spherical coordinate system depend on the position. When a particle moves, the directions of the basis vectors change as well. This is a different way of saying that the spherical coordinate system is not a Cartesian system where the orientation of the coordinate axes is independent of the position. When computing the acceleration in such a system additional terms appear that account for the fact that the coordinate system is curvilinear. The results of Section 4.1 contain all the ingredients we need.

Let us follow a particle moving over a sphere. The position vector **r** has an obvious expansion in spherical coordinates:

$$\mathbf{r} = r\hat{\mathbf{r}}. \tag{4.21}$$

The velocity is obtained by taking the time-derivative of this expression. However, the unit vector $\hat{\mathbf{r}}$ is a function of the angles θ and φ, see equation (4.7). This means that when we take the time-derivative of (4.21) to obtain the velocity we also need to differentiate $\hat{\mathbf{r}}$ with time. Note that this is not the case with the Cartesian expression $\mathbf{r} = x\hat{\mathbf{x}} + y\hat{\mathbf{y}} + z\hat{\mathbf{z}}$ because the unit vectors $\hat{\mathbf{x}}$, $\hat{\mathbf{y}}$, and $\hat{\mathbf{z}}$ are constant, hence they do not change when the particle moves and they thus have a vanishing time-derivative.

As an example, let us compute the time-derivative of $\hat{\mathbf{r}}$. This vector is a function of θ and φ, and these angles both change with time as the particle moves. Using the chain rule it thus follows that:

$$\frac{d\hat{\mathbf{r}}}{dt} = \frac{d\hat{\mathbf{r}}(\theta, \varphi)}{dt} = \frac{d\theta}{dt}\frac{\partial \hat{\mathbf{r}}}{\partial \theta} + \frac{d\varphi}{dt}\frac{\partial \hat{\mathbf{r}}}{\partial \varphi}. \tag{4.22}$$

The derivatives $\partial\hat{\mathbf{r}}/\partial\theta$ and $\partial\hat{\mathbf{r}}/\partial\varphi$ can be eliminated with (4.12).

Problem a Use expressions (4.12) to eliminate the derivatives $\partial\hat{\mathbf{r}}/\partial\theta$ and $\partial\hat{\mathbf{r}}/\partial\varphi$ and carry out a similar analysis for the time-derivatives of the unit vectors $\hat{\boldsymbol{\theta}}$

and $\hat{\boldsymbol{\varphi}}$ to show that:

$$\left.\begin{aligned} \frac{d\hat{\mathbf{r}}}{dt} &= \dot{\theta}\hat{\boldsymbol{\theta}} + \sin\theta\ \dot{\varphi}\ \hat{\boldsymbol{\varphi}}, \\ \frac{d\hat{\boldsymbol{\theta}}}{dt} &= -\dot{\theta}\hat{\mathbf{r}} + \cos\theta\ \dot{\varphi}\ \hat{\boldsymbol{\varphi}}, \\ \frac{d\hat{\boldsymbol{\varphi}}}{dt} &= -\sin\theta\ \dot{\varphi}\ \hat{\mathbf{r}} - \cos\theta\ \dot{\varphi}\ \hat{\boldsymbol{\theta}}. \end{aligned}\right\} \tag{4.23}$$

In these and other expressions in this section a dot is used to denote the time-derivative: $\dot{F} \equiv dF/dt$.

Problem b Use (4.21), the first line of (4.23) and the definition $\mathbf{v} = d\mathbf{r}/dt$ to show that in spherical coordinates:

$$\mathbf{v} = \dot{r}\hat{\mathbf{r}} + r\dot{\theta}\hat{\boldsymbol{\theta}} + r\sin\theta\ \dot{\varphi}\ \hat{\boldsymbol{\varphi}}. \tag{4.24}$$

In spherical coordinates the components of the velocity are thus given by:

$$\left.\begin{aligned} v_r &= \dot{r}, \\ v_\theta &= r\dot{\theta}, \\ v_\varphi &= r\sin\theta\ \dot{\varphi}. \end{aligned}\right\} \tag{4.25}$$

This result can be interpreted geometrically. As an example, let us consider the radial component of the velocity, see Figure 4.2. To obtain the radial component of the velocity we keep the angles θ and φ fixed and let the radius $r(t)$ change to $r(t + \Delta t)$ over a time Δt. The particle has moved a distance $r(t + \Delta t) - r(t) = (dr/dt)/\Delta t$ in a time Δt, so that the radial component of the velocity is given by $v_r = dr/dt = \dot{r}$. This is the result given by the first line of (4.25).

Problem c Use similar geometric arguments to explain the form of the velocity components v_θ and v_φ given in (4.25).

Problem d We are now in the position to compute the acceleration in spherical coordinates. To do this differentiate (4.24) with respect to time and use expression (4.23) to eliminate the time-derivatives of the basis vectors. Use this to show that the acceleration $\mathbf{a}$ is given by:

$$\begin{aligned} \mathbf{a} = {} & \left(\dot{v}_r - \dot{\theta}v_\theta - \sin\theta\ \dot{\varphi}v_\varphi\right)\hat{\mathbf{r}} \\ & + \left(\dot{v}_\theta + \dot{\theta}v_r - \cos\theta\ \dot{\varphi}v_\varphi\right)\hat{\boldsymbol{\theta}} \\ & + \left(\dot{v}_\varphi + \sin\theta\ \dot{\varphi}v_r + \cos\theta\ \dot{\varphi}v_\theta\right)\hat{\boldsymbol{\varphi}}. \end{aligned} \tag{4.26}$$

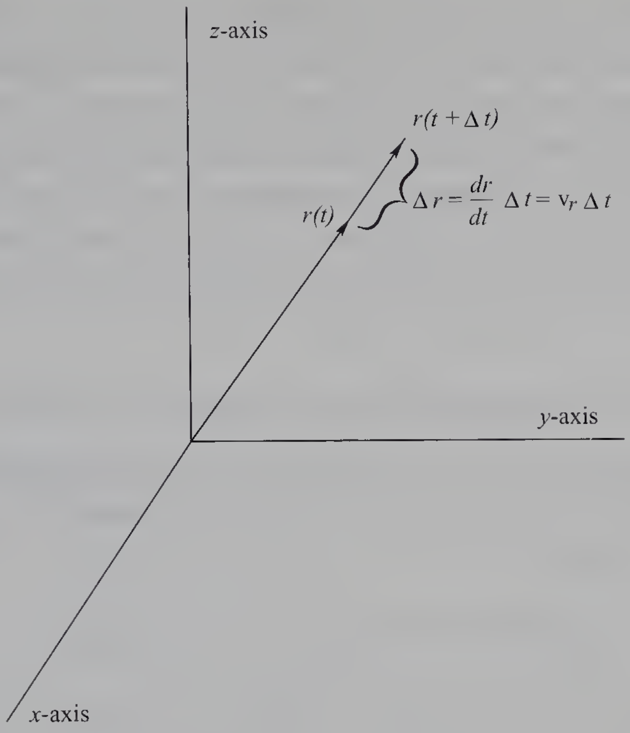

Fig. 4.2 Definition of the geometric variables used to derive the radial component of the velocity.

Problem e This expression is not quite satisfactory because it contains both the components of the velocity as well as the time-derivatives $\dot{\theta}$ and $\dot{\varphi}$ of the angles. Eliminate the time-derivatives with respect to the angles in favor of the components of the velocity using expressions (4.25) to show that the components of the acceleration in spherical coordinates are given by:

$$\left.\begin{aligned} a_r &= \dot{v}_r - \frac{v_\theta^2 + v_\varphi^2}{r}, \\ a_\theta &= \dot{v}_\theta + \frac{v_r v_\theta}{r} - \frac{v_\varphi^2}{r \tan\theta}, \\ a_\varphi &= \dot{v}_\varphi + \frac{v_r v_\varphi}{r} + \frac{v_\theta v_\varphi}{r \tan\theta}. \end{aligned}\right\} \tag{4.27}$$

It thus follows that the components of the acceleration in a spherical coordinate system are not simply the time-derivatives of the components of the velocity in that system. The reason for this is that the spherical coordinate system uses basis vectors that change when the particle moves. Expressions (4.27) play a crucial role in meteorology and oceanography where one describes the motion of the atmosphere or ocean [50]. Of course, in that application one should account for the Earth's rotation as well so that terms accounting for the Coriolis force and the centrifugal force need to be added, see Section 13.3. It should also be noted that the analysis of this section has been oversimplified when applied to the ocean or atmosphere. The reason for this is that the time-derivative of a moving parcel of air or water contains the explicit time-derivative $\partial/\partial t$ as well as a term $\mathbf{v} \cdot \boldsymbol{\nabla}$ that accounts for the fact that the properties change because the parcel moves to another location in space. The difference between the partial derivative $\partial/\partial t$ and the total derivative d/dt is treated in Section 5.5; this distinction has not been taken into account in the analysis in this section. A complete treatment is given by Holton [50].

4.4 Volume integration in spherical coordinates

Carrying out a volume integration in Cartesian coordinates involves multiplying the function to be integrated by an infinitesimal volume element $dxdydz$ and integrating over all volume elements:

$$\iiint F dV = \iiint F(x, y, z)\, dxdydz.$$

Although this seems to be a simple procedure, it can be quite complex when the function F depends in a complex way on the coordinates (x, y, z) or when the limits of integration are not simple functions of x, y, and z.

Problem a Compute the volume of a sphere of radius R by taking $F = 1$ and integrating the volume integral in Cartesian coordinates over the volume of the sphere. Show first that in Cartesian coordinates the volume of the sphere can be written as

$$volume = \int_{-R}^{R} \int_{-\sqrt{R^2-x^2}}^{\sqrt{R^2-x^2}} \int_{-\sqrt{R^2-x^2-y^2}}^{\sqrt{R^2-x^2-y^2}} dzdydx, \tag{4.28}$$

and then carry out the integrations.

After carrying out this exercise you have probably become convinced that using Cartesian coordinates is not the most efficient way to derive that the volume of a sphere with radius R is given by $4\pi R^3/3$. Using spherical coordinates appears to be the way to go, but for this one needs to be able to express an infinitesimal volume element dV in spherical coordinates. In doing this we will use that the volume spanned by three vectors **a**, **b**, and **c** is given by

$$volume = \det(\ \mathbf{a},\ \mathbf{b},\ \mathbf{c}\ ) = \begin{vmatrix} a_x & b_x & c_x \\ a_y & b_y & c_y \\ a_z & b_z & c_z \end{vmatrix}, \tag{4.29}$$

see reference [19]. If we change the spherical coordinate θ by an increment $d\theta$, the position vector changes from $\mathbf{r}(r, \theta, \varphi)$ to $\mathbf{r}(r, \theta + d\theta, \varphi)$, and this corresponds to a change $\mathbf{r}(r, \theta + d\theta, \varphi) - \mathbf{r}(r, \theta, \varphi) = \partial\mathbf{r}/\partial\theta \, d\theta$ in the position vector. Using the same reasoning for the variation of the position vector with r and φ it follows that the infinitesimal volume dV corresponding to increments dr, $d\theta$, and $d\varphi$ is given by

$$dV = \det\left(\frac{\partial\mathbf{r}}{\partial r}dr, \ \frac{\partial\mathbf{r}}{\partial\theta}d\theta, \ \frac{\partial\mathbf{r}}{\partial\varphi}d\varphi\right). \tag{4.30}$$

Problem b Show that this can be written as:

$$dV = \underbrace{\begin{vmatrix} \frac{\partial x}{\partial r} & \frac{\partial x}{\partial\theta} & \frac{\partial x}{\partial\varphi} \\ \frac{\partial y}{\partial r} & \frac{\partial y}{\partial\theta} & \frac{\partial y}{\partial\varphi} \\ \frac{\partial z}{\partial r} & \frac{\partial z}{\partial\theta} & \frac{\partial z}{\partial\varphi} \end{vmatrix}}_{J} dr d\theta d\varphi = J dr d\theta d\varphi. \tag{4.31}$$

The determinant J is called the *Jacobian*, which is also sometimes written as:

$$J = \frac{\partial(x, y, z)}{\partial(r, \theta, \varphi)}. \tag{4.32}$$

It should be kept in mind that this is nothing more than a new notation for the determinant in (4.31).

Problem c Use expressions (4.3) and (4.31) to show that

$$J = r^2 \sin\theta. \tag{4.33}$$

Note that the Jacobian J in (4.33) is the product of the scale factors defined in equation (4.8): $J = h_r h_\theta h_\varphi$. This is not a coincidence; in general the scale factors contain all the information needed to compute the Jacobian for an orthogonal curvilinear coordinate system, see Butkov [24] for details.

Problem d A volume element dV is thus given in spherical coordinates by $dV = r^2 \sin\theta \, dr d\theta d\varphi$. Consider the volume element dV in Figure 4.3 that is defined by infinitesimal increments dr, $d\theta$, and $d\varphi$. Give an alternative derivation of this expression for dV that is based on geometric arguments only.

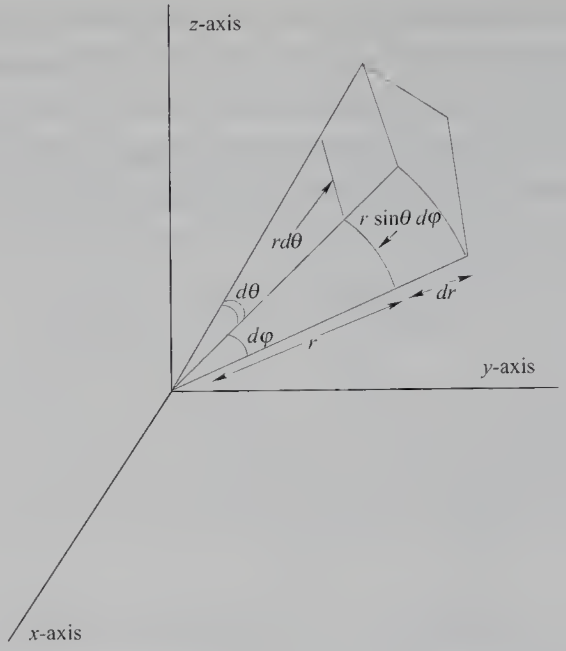

Fig. 4.3 Definition of the geometric variables for an infinitesimal volume element dV.

In some applications one wants to integrate over the surface of a sphere rather than over a volume. For example, if one wants to compute the cooling of the Earth, one needs to integrate the heat flow over the Earth's surface. The treatment used for deriving the volume integral in spherical coordinates can also be used to derive the surface integral. A key element in the analysis is that the surface spanned by two vectors **a** and **b** is given by $|\mathbf{a} \times \mathbf{b}|$. Again, an increment $d\theta$ of the angle θ corresponds to a change $(\partial \mathbf{r}/\partial\theta)\, d\theta$ of the position vector. A similar result holds when the angle φ is changed.

Problem e Use these results to show that the surface element dS corresponding to infinitesimal changes $d\theta$ and $d\varphi$ is given by

$$dS = \left| \frac{\partial \mathbf{r}}{\partial \theta} \times \frac{\partial \mathbf{r}}{\partial \varphi} \right| d\theta d\varphi. \tag{4.34}$$

In deriving this you can use that the area spanned by two vectors **v** and **w** is given by $A = |\mathbf{v}|\,|\mathbf{w}| \sin\psi = |\mathbf{v} \times \mathbf{w}|$, where ψ is the angle between these vectors.

Problem f Use expression (4.3) to compute the vectors in the cross product and use this to derive that

$$dS = r^2 \sin\theta \; d\theta d\varphi. \tag{4.35}$$

Problem g Using the geometric variables in Figure 4.3 give an alternative derivation of this expression for a surface element that is based on geometric arguments only.

Problem h Compute the volume of a sphere with radius R using spherical coordinates. Pay special attention to the range of integration for the angles θ and φ, see Section 4.1.

4.5 Cylindrical coordinates

Cylindrical coordinates are useful in problems that exhibit cylindrical symmetry rather than spherical symmetry. An example is the generation of water waves when a stone is thrown into a pond, or more importantly when an earthquake excites a tsunami in the ocean. In cylindrical coordinates a point is specified by giving its distance $r = \sqrt{x^2 + y^2}$ to the z-axis, the angle φ, and the z-coordinate, see Figure 4.4 for the definition of the variables. All the results we need could be derived using an analysis like those shown in the previous sections. However, in such an approach we would do a large amount of unnecessary work. The key is to realize that at the equator of a spherical coordinate system (i.e. at the locations where $\theta = \pi/2$) the spherical coordinate system and the cylindrical coordinate system are identical, see Figure 4.5. An inspection of this figure shows that all results obtained for spherical coordinates can be used for cylindrical coordinates by making the

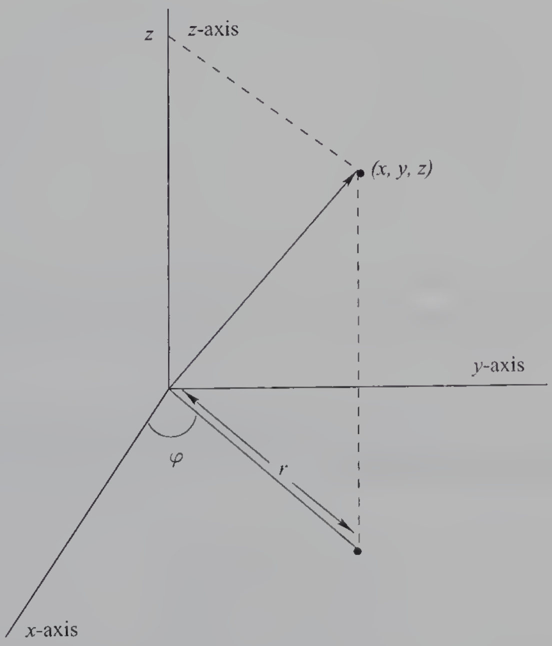

Fig. 4.4 Definition of the geometric variables used in cylindrical coordinates.

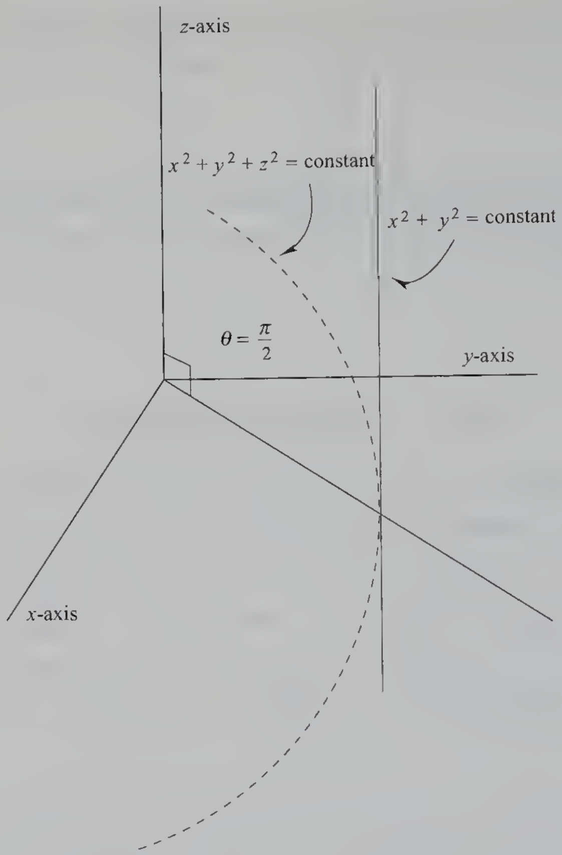

Fig. 4.5 At the equator the spherical coordinate system has the same properties as a system of cylindrical coordinates.

following substitutions:

$$\left.\begin{array}{l} r=\sqrt{x^2+y^2+z^2} \rightarrow \sqrt{x^2+y^2}, \\ \theta \rightarrow \pi/2, \\ \hat{\boldsymbol{\theta}} \rightarrow -\hat{\mathbf{z}}, \\ r d\theta \rightarrow -dz. \end{array}\right\} \tag{4.36}$$

Problem a Convince yourself of this. To derive the third line consider the unit vectors pointing in the direction of increasing values of θ and z at the equator.

Problem b Use the results of the previous sections and the substitutions (4.36) to show the following properties for a system of cylindrical coordinates:

$$\left.\begin{array}{l} x=r\cos\varphi, \\ y=r\sin\varphi, \\ z=z. \end{array}\right\} \tag{4.37}$$

$$\hat{\mathbf{r}} = \begin{pmatrix} \cos\varphi \\ \sin\varphi \\ 0 \end{pmatrix}, \quad \hat{\boldsymbol{\varphi}} = \begin{pmatrix} -\sin\varphi \\ \cos\varphi \\ 0 \end{pmatrix}, \quad \hat{\mathbf{z}} = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}, \tag{4.38}$$

$$dV = r dr d\varphi dz, \tag{4.39}$$

$$dS = r dz d\varphi. \tag{4.40}$$

Problem c Derive these properties directly using geometric arguments.

5

Gradient

In this chapter the gradient of a function is introduced. The gradient plays an important role in the differentiation and integration of functions in more than one dimension (Section 5.3). Newton's law is derived in Section 5.4 from the concept of energy conservation. As a by-product of this derivation it follows that the force is the (negative) gradient of the potential energy. The gradient plays a crucial role in the distinction between the partial time derivatives and the total time derivatives. This resulting distinction between an Eulerian and a Lagrangian formulation of problems involving fluid flow is shown in Section 5.5. In Section 5.6 expressions in spherical coordinates and cylindrical coordinates for the gradient are derived.

5.1 Properties of the gradient vector

Let us consider a function f that depends on the variables x and y in a plane. We want to describe how this function changes when we move from point A in the plane to point C as shown in Figure 5.1. The resulting change in the function f is denoted by $\delta f = f_C - f_A$, where f_A denotes, for example, the function f at point A. It follows from Figure 5.1 that

$$\left.\begin{aligned} f_A &= f(x, y), \\ f_B &= f(x + \delta x, y), \\ f_C &= f(x + \delta x, y + \delta y). \end{aligned}\right\} \tag{5.1}$$

It follows by adding and subtracting f_B that

$$\delta f = f_C - f_A = f_B - f_A + f_C - f_B. \tag{5.2}$$

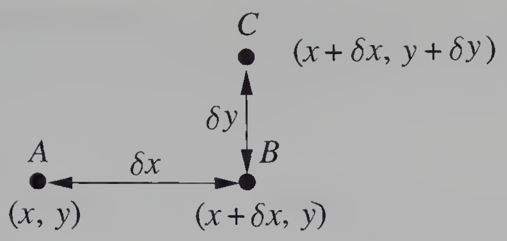

Fig. 5.1 Definition of the points A, B, and C.

Problem a Use (5.1) and expression (3.18) to derive that for small values of δx and δy:

$$\left.\begin{aligned} f_B - f_A &= \frac{\partial f}{\partial x}(x, y)\,\delta x, \\ f_C - f_B &= \frac{\partial f}{\partial y}(x+\delta x, y)\,\delta y. \end{aligned}\right\} \tag{5.3}$$

Problem b Insert this result in (5.2) and derive that to leading order in δx and δy the result can be written as:

$$\delta f = \frac{\partial f}{\partial x}(x, y)\,\delta x + \frac{\partial f}{\partial y}(x, y)\,\delta y. \tag{5.4}$$

Note that $\partial f/\partial y$ is evaluated at $(x+\delta x, y)$. However, using a Taylor expansion of $\partial f/\partial y$ we find that

$$\frac{\partial f}{\partial y}(x+\delta x, y)\,\delta y = \frac{\partial f}{\partial y}(x, y)\,\delta y + \frac{\partial^2 f}{\partial x \partial y}(x, y)\,\delta x \delta y. \tag{5.5}$$

The last term is of second order in δx and δy and can be ignored when these quantities are small.

Equation (5.4) has the same form as the inner product between two vectors **a** and **b** in two dimensions: $(\mathbf{a}\cdot\mathbf{b}) = a_x b_x + a_y b_y$. Let us define $\delta\mathbf{r}$ as the difference in the location vector of the points C and A: $\delta\mathbf{r} \equiv \mathbf{r}_C - \mathbf{r}_A$. This vector has components δx and δy so that in two dimensions:

$$\delta\mathbf{r} = \begin{pmatrix} \delta x \\ \delta y \end{pmatrix}. \tag{5.6}$$

Similarly we define a vector that contains as the x-component the partial x-derivative of the function f and as the y-component the partial y-derivative:

$$\nabla f \equiv \begin{pmatrix} \partial f/\partial x \\ \partial f/\partial y \end{pmatrix}. \tag{5.7}$$

This vector is called the *gradient of f*. It is customary to denote the gradient of a function f by the symbol ∇f, but another notation you may find in the literature is grad f.

Problem c Show that the increment δf of expression (5.4) can be expressed as the inner product of two vectors:

$$\delta f = (\nabla f \cdot \delta \mathbf{r}). \tag{5.8}$$

This expression shows directly why the gradient is such a useful vector. Once we know the gradient ∇f, we can use (5.8) to compute the change in the function f when we change the point of evaluation with an arbitrary step $\delta \mathbf{r}$. Note that this expression holds for *any* direction in which we can take this step. It should be noted that (5.8) is based on the first order Taylor expansions in expression (5.3). This means that (5.8) only holds in the limit $\delta \mathbf{r} \rightarrow 0$. However, this expression is still extremely useful because it forms the basis of the rules for differentiation and integration in more than one dimension. We will return to this issue in Section 5.3.

Problem d The derivation up to this point has been for two space dimensions only. Repeat this derivation for three space dimensions and derive that (5.8) still holds when the gradient in three dimensions is defined as

$$\nabla f \equiv \begin{pmatrix} \partial f/\partial x \\ \partial f/\partial y \\ \partial f/\partial z \end{pmatrix} \quad \text{and} \quad \delta \mathbf{r} \equiv \begin{pmatrix} \delta x \\ \delta y \\ \delta z \end{pmatrix}. \tag{5.9}$$

Problem e Compute ∇f when $f(x, y, z) = x\, e^{-y} \sin z$.

The gradient is a vector and it therefore has a direction and a magnitude. The direction of the gradient can be obtained from expression (5.8). Let us consider a change $\delta \mathbf{r}$ such that the function f does not change in this direction. This means that $\delta \mathbf{r}$ is chosen in such a way that the displacement is along a surface where $f = const.$ as depicted in Figure 5.2. For such a change $\delta \mathbf{r}$, the corresponding change δf is by definition equal to zero, hence $(\nabla f \cdot \delta \mathbf{r}) = 0$.

Problem f Use the properties of the inner product of two vectors to show that this identity implies that the vectors ∇f and $\delta \mathbf{r}$ are for this special perturbation perpendicular: $\nabla f \perp \delta \mathbf{r}$.

This last identity of course only holds when the step $\delta \mathbf{r}$ is taken along the surface where the function f is constant. However, it does imply that the gradient vector is

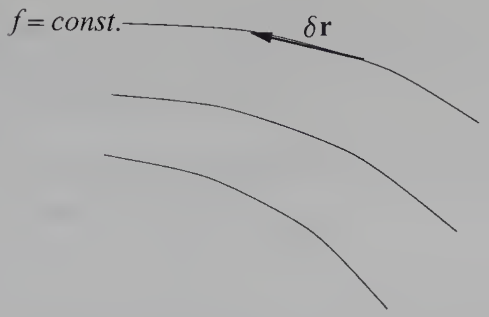

Fig. 5.2 Contour lines defined by the condition $f = const.$ and a perturbation $\delta\mathbf{r}$ in the position vector along a contour line.

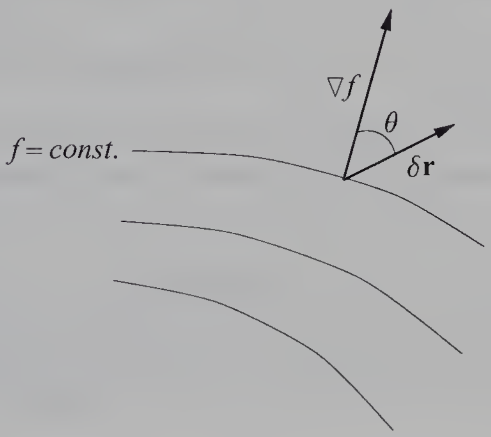

Fig. 5.3 Contour lines defined by the condition $f = const.$ and a perturbation $\delta\mathbf{r}$ in the position vector in an arbitrary direction.

perpendicular to the surface $f = const.$ Now that we know this, the gradient can still point in two directions because one can move in two directions perpendicular to a surface. Consider Figure 5.3 where we consider a step $\delta\mathbf{r}$ in an arbitrary direction.

Problem g Use (5.8) to show that

$$\delta f = |\nabla f|\,|\delta\mathbf{r}|\cos\theta, \tag{5.10}$$

where θ is the angle between ∇f and $\delta\mathbf{r}$.

The change δf is largest when the vectors ∇f and $\delta\mathbf{r}$ point in the same direction because in that case $\cos\theta$ is equal to its maximum value of 1, that is when $\theta = 0$. We also know that δf increases most rapidly when $\delta\mathbf{r}$ is directed in such a way that one moves from smaller values of f towards larger values of f. Since δf is largest when $\theta = 0$ this means the gradient also points from small values of f towards high values of f.

The magnitude of the gradient vector can also be obtained from (5.10). Let us consider a step $\delta \mathbf{r}$ in the direction of the gradient vector, hence a step in the direction of increasing values of f. In that case the vectors ∇f and $\delta \mathbf{r}$ are parallel, and $\cos\theta = 1$. This means that (5.10) implies that

$$|\nabla f| = \frac{\delta f}{|\delta \mathbf{r}|}, \tag{5.11}$$

where $\delta \mathbf{r}$ is a step in the direction of increasing values of f. Summarizing this we obtain the following properties of the gradient vector ∇f:

1. The gradient of a function f is a vector that is perpendicular to the surface $f = const.$
2. The gradient points in the direction of increasing values of f.
3. The magnitude of the gradient is the change δf in the function in the direction of the largest increase divided by the distance $|\delta \mathbf{r}|$ in that direction.

5.2 Pressure force

An example of a two-dimensional function is shown in Figure 5.4 where the atmospheric pressure on the Earth's surface is shown in units of millibars.

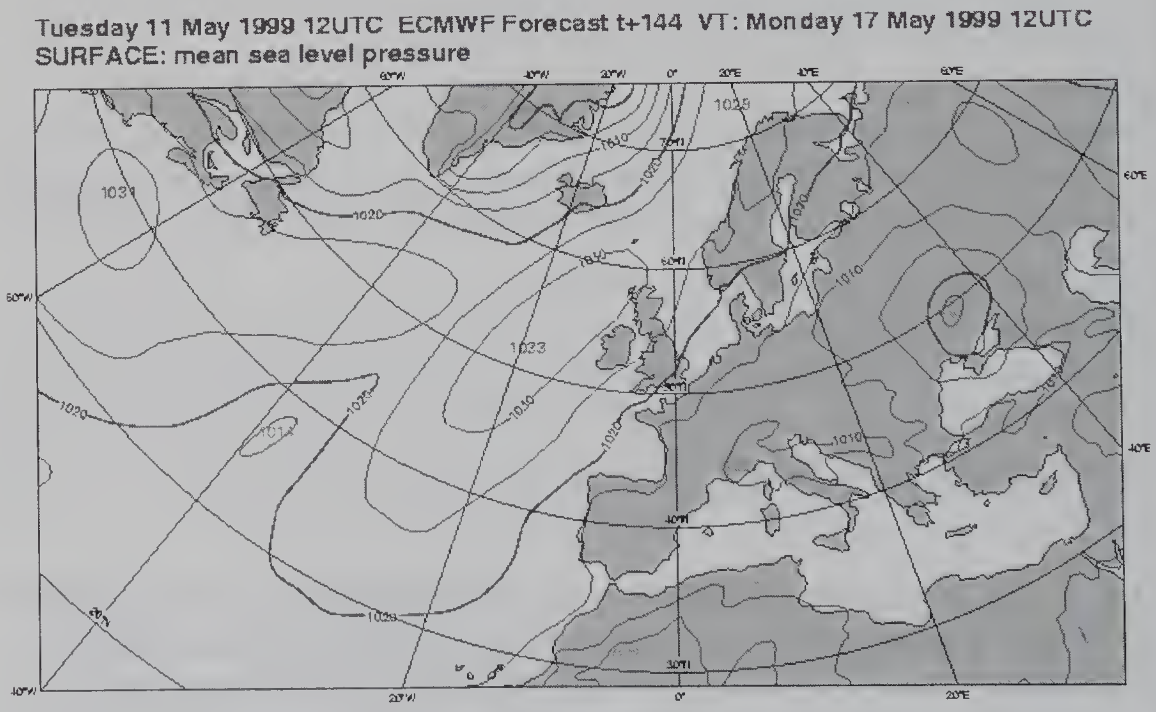
Tuesday 11 May 1999 12UTC ECMWF Forecast t+144 VT: Monday 17 May 1999 12UTC
SURFACE: mean sea level pressure
1031
1020
1029
1010
1023
1014
1010
1028

Fig. 5.4 Contour map of the pressure at the Earth's surface on 17 May 1999 in units of millibars. (Figure courtesy of the European Centre for Medium-Range Weather Forecasts (ECMWF).)

Problem a Draw the gradient vector of the pressure p in a number of locations on this map. When doing so draw larger arrows where the gradient is greater.

Problem b What is the approximate gradient $|\nabla p|$ over Ireland? Is the gradient vector larger over Ireland than over Iceland?

There is a good reason why a map of the pressure is used here to illustrate the concept of the gradient. When the air pressure is not constant, a parcel of air will experience a force that pushes it from a region of high pressure towards a region of lower air pressure. This force increases when the pressure varies more rapidly with distance. This can be described by the pressure force $\mathbf{F}_p$ that is given by:

$$\mathbf{F}_p = -\nabla p. \tag{5.12}$$

Problem c The pressure in the weather map of Figure 5.4 is shown in units of millibars. If you have worked out problem b properly you will have deduced that the units of the pressure force are in millibars per kilometre. A pressure is a force per unit area because the pressure times the area gives the total force that acts on this area. Deduce from this result that the pressure force has the dimensions *force/volume*.

Problem d Use expression (5.12) to show that the pressure force points from regions of high pressure towards regions of lower pressure. Draw the direction of the pressure force at the high-pressure area in Figure 5.4 over Ireland.

You may have noted that we have not really derived expression (5.12) for the pressure force. Sometimes only physical arguments are used to state a physical law. However, one often needs to verify that the arguments employed have a proper mathematical basis. It is possible to derive the pressure force from the fact that the force that acts in a gas or a fluid on a hypothetical surface within that medium is perpendicular to that surface and its strength is given by the pressure times the surface area.

Problem e Consider a small volume element with lengths δx, δy, and δz in the three coordinate directions as shown in Figure 5.5. Let us consider the net force in the x-direction. Use the concept of pressure as described above to show that the force acting on the left-hand plane of the volume is given by $p(x, y, z)\delta y \delta z\ \hat{\mathbf{x}}$ and that the force acting on the right-hand plane is given by $-p(x + \delta x, y, z)\delta y \delta z\ \hat{\mathbf{x}}$, where $\hat{\mathbf{x}}$ is the unit vector in the x-direction. Explain the minus sign in the second term.

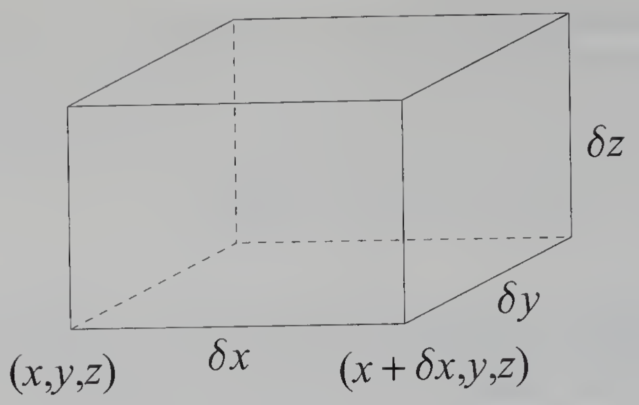

Fig. 5.5 Definition of geometric variables for the derivation of the pressure force.

The top, bottom, front, and back surfaces do not contribute to the x-component of the pressure force because the pressure force acting on these surfaces is directed in the y- or z-direction. This implies that the x-component of the total force is given by

$$f_x = -\left[p(x+\delta x, y, z) - p(x, y, z)\right] \delta y \delta z \ \hat{\mathbf{x}}. \tag{5.13}$$

Problem f Use expression (3.18) to show that this can also be written as

$$f_x = -\frac{\partial p}{\partial x} \delta V \ \hat{\mathbf{x}}, \tag{5.14}$$

where the volume δV is equal to $\delta x \delta y \delta z$.

Problem g Apply the same reasoning to obtain the y- and z-components of the force and show that the net force felt by the volume is given by

$$\mathbf{f} = -\nabla p \ \delta V. \tag{5.15}$$

You may have been puzzled by the fact that in problem c you deduced that the pressure force $\mathbf{F}_p$ has the dimensions $force/volume$. This means it is not really a force. In fact, we see in (5.15) that the net force that acts on the volume is given by $-\nabla p \delta V$. You have to keep in mind that the volume δV is not a physical entity, instead it is a mathematical volume used in our reasoning. The net force is proportional to δV, and since this volume is physically meaningless the net force also has no physical meaning. However, when ρ is the mass-density of the gas or fluid then $\delta m = \rho \ \delta V$ is the mass of the volume. This means that both the mass of the volume and the net force that acts on the volume are proportional to δV. When we apply Newton's law $\delta m \mathbf{a} = \mathbf{f}$ to this volume, we can divide both the left-hand side and the right-hand side by the arbitrary volume δV and Newton's law then takes the form $\rho \mathbf{a} = \mathbf{F}_p$, with the pressure force given by (5.12).

This means that the pressure force $\mathbf{F}_p$ should be seen as a force per unit volume, just as the density ρ is the mass per unit volume. In fluid mechanics one always works with physical quantities per unit volume, for the simple reason that a gas or fluid is not composed of physical small volumes. We return to a more rigorous treatment of Newton's law in fluid dynamics in Section 11.3 where the conservation of momentum in a continuous medium is treated.

The pressure force is one of the most important forces in fluid mechanics, meteorology, and oceanography because it is the variation in the pressure that underlies the motion of fluids and gases. Physically it states the simple fact that a gas or fluid is pushed away from regions of higher pressure to regions of lower pressure.

5.3 Differentiation and integration

The results of Section 5.1 hold for infinitesimal changes $\delta \mathbf{r}$ only. However, a change over a finite distance can be thought of as being built up from many infinitesimal steps. Let us consider two points A and B that may be far apart as shown in Figure 5.6. Between the points A and B we can insert a large number of points $P_1, P_2, \ldots, P_N$. The difference in the function values at the points A and B can then be written as

$$f_B - f_A = (f_B - f_N) + (f_N - f_{N-1}) + \cdots + (f_2 - f_1) + (f_1 - f_A), \quad (5.16)$$

where f_j denotes the function evaluated at point P_j. What we are really doing is dividing the large distance from A to B into infinitesimally smaller intervals. If we take enough of these subintervals we can apply (5.8) to each of these sub-intervals.

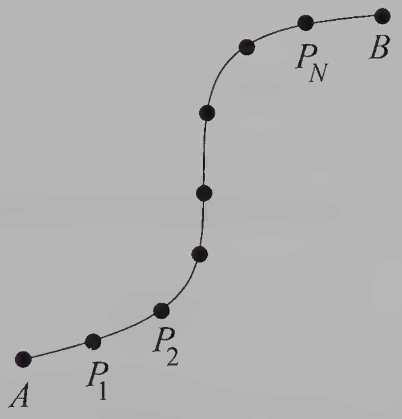

Fig. 5.6 The definition of the points A and B at the end of an interval and the intermediary points $P_1, P_2, \ldots, P_N$.

Problem a Show that

$$f_B - f_A = \sum (\nabla f \cdot \delta\mathbf{r}), \tag{5.17}$$

where the sum is over the sub-intervals used in (5.16) and where $\delta\mathbf{r}$ is the increment in the position vector in each interval.

The analysis in this section holds only in the limit $N \to \infty$ where the interval is divided into infinitesimal intervals. In that case the summation in (5.17) is replaced by an integration and the notation $d\mathbf{r}$ is used rather than $\delta\mathbf{r}$:

$$f_B - f_A = \int_A^B (\nabla f \cdot d\mathbf{r}). \tag{5.18}$$

This expression is extremely useful because it makes it possible to compute the change in a function between two points once one knows the gradient of that function. It should be noted that (5.18) holds for *any* path that joins the points A and B. This can easily be seen from the derivation of this section, because the points $P_1, P_2, \ldots, P_N$ can be arbitrarily chosen as long as they form a continuous path that joins the points A and B. This property can sometimes be exploited by choosing the path for which the calculation of the integral is easiest.

Problem b Use these results to show that the line integral of ∇f along any closed contour is equal to zero:

$$\oint (\nabla f \cdot d\mathbf{r}) = 0. \tag{5.19}$$

The gradient can also be used to determine the derivative of a function in a given direction. The *directional derivative* df/ds of the function in the direction of a unit vector $\hat{\mathbf{n}}$ is defined as the change of the function in the direction of $\hat{\mathbf{n}}$ normalized per unit distance:

$$\frac{df}{ds}(\mathbf{r}) = \lim_{\delta s \to 0} \frac{f(\mathbf{r} + \hat{\mathbf{n}}\delta s) - f(\mathbf{r})}{\delta s}. \tag{5.20}$$

Problem c Use (5.8) in the numerator of this expression to derive that

$$\frac{df}{ds}(\mathbf{r}) = (\hat{\mathbf{n}} \cdot \nabla f). \tag{5.21}$$

This expression is important because it allows us to compute the derivative of a function in *any* arbitrary direction once the gradient is known.

Equations (5.18) and (5.21) generalize the rules for the integration and differentiation of functions of one variable to more space dimensions.

Problem d To see this, let the function f depend on the variable x only and let the points A and B be located on the x-axis. Let $\hat{\mathbf{x}}$ denote the unit vector in the direction of the x-axis. Use the relation $d\mathbf{r} = \hat{\mathbf{x}}dx$ to show that in that case (5.18) reduces to the well-known integration rule of a function that depends only on one variable:

$$f_B - f_A = \int_A^B \frac{\partial f}{\partial x} dx. \tag{5.22}$$

Problem e Let the vector $\hat{\mathbf{n}}$ be directed along the x-axis so that $\hat{\mathbf{n}} = \hat{\mathbf{x}}$. Show that in that case the directional derivative along the x-axis in (5.21) is given by $\partial f/\partial x$.

5.4 Newton's law from energy conservation

In classical mechanics one can start from Newton's law and derive that the total energy of a mechanical system without friction is conserved. You may have discovered that in physics there is often no proof of the basic physical laws. For example, there is no "proof" of Newton's law. Its use is justified by the observation that it describes the motion of the Sun, Moon, and planets very accurately. To a certain extent, it is arbitrary which physical law one uses as a starting point. In this section we use the concept of energy conservation as a starting point and then derive Newton's law.

Let us consider a mechanical system without friction. The total energy E is the sum of the kinetic energy $\frac{1}{2}mv^2$ and the potential energy $V(\mathbf{r})$:

$$\frac{1}{2}mv^2 + V(\mathbf{r}) = E. \tag{5.23}$$

We assume that in such a system the total energy E is conserved; this means that the time derivative of this quantity is equal to zero

$$\frac{dE}{dt} = 0. \tag{5.24}$$

In order to derive Newton's law from this expression we need to take the time derivative of both the kinetic and the potential energy.

Problem a Show that the time derivative of the kinetic energy is given by

$$\frac{d}{dt}\left(\frac{1}{2}mv^2\right) = m\left(\mathbf{v}\cdot\frac{d\mathbf{v}}{dt}\right). \tag{5.25}$$

Hint: write $v^2 = v_x^2 + v_y^2 + v_z^2$ and differentiate each term.

We also need to compute the time derivative of the potential energy. At first you may be tempted to conclude that this time derivative is equal to zero because the potential depends on the position **r** but not on time. However, as the particle moves through space it is at different positions at different times. Therefore, the position of the particle is a function of time: $\mathbf{r} = \mathbf{r}(t)$. For this reason, the potential should be written as $V(\mathbf{r}(t))$. The time derivative of the potential then follows from the usual rule for taking a derivative:

$$\frac{dV(\mathbf{r})}{dt} = \lim_{\delta t\to 0}\frac{V(\mathbf{r}(t+\delta t)) - V(\mathbf{r}(t))}{\delta t}. \tag{5.26}$$

Problem b Use (5.8) to show that this time derivative can be written as

$$\frac{dV(\mathbf{r})}{dt} = \lim_{\delta t\to 0}\frac{(\nabla V\cdot\delta\mathbf{r})}{\delta t}, \tag{5.27}$$

with $\delta\mathbf{r} = \mathbf{r}(t+\delta t) - \mathbf{r}(t)$.

Problem c Use the definition of the velocity $\mathbf{v} = \lim_{\delta t\to 0}\delta\mathbf{r}/\delta t$ to show that

$$\frac{dV(\mathbf{r})}{dt} = (\mathbf{v}\cdot\nabla V). \tag{5.28}$$

Problem d Use these results to show that the law of energy conservation (5.24) can be written as

$$\mathbf{v}\cdot\left(m\frac{d\mathbf{v}}{dt} + \nabla V\right) = 0. \tag{5.29}$$

Since **v** is arbitrary, (5.29) must hold for any velocity vector **v**. This can only be the case when the term in brackets is equal to zero.

Problem e Show that this implies that

$$m\frac{d\mathbf{v}}{dt} = \mathbf{F}, \tag{5.30}$$

with

$$\mathbf{F} = -\nabla V. \tag{5.31}$$

Equation (5.30) is Newton's law; we have derived it here from the requirement of energy conservation. As a by-product we have shown that the force in Newton's law is equal to $-\nabla V$. Note that the force (5.31) and the pressure force (5.12) share the common property that they follow from the gradient of a scalar function. You may find this analogy useful. We have argued in Section 5.2 that the pressure force pushes air from a region of higher pressure to a region of lower pressure. By the same reasoning the force $\mathbf{F} = -\nabla V$ pushes a particle from regions of higher potential energy to regions of lower potential energy. A simple example of this is gravity. The gravitational potential energy increases with the distance to the attracting body. This means that according to the reasoning given here gravitation forces a particle closer to the attracting body. This is an observation that may be obvious to you when you see an apple fall from a tree. However, you may want to keep the analogy of the pressure force and the force associated with a general potential energy in mind because it helps to understand the concept of potential energy.

With all the elements we have assembled here, we can now derive the concept of power. This quantity is defined as the energy delivered to the particle per unit time. This means that the power is defined as the time derivative of the kinetic energy.

Problem f Use (5.24), (5.28), and (5.31) to derive that the power is given by

$$\frac{d}{dt}\left(\frac{1}{2}mv^2\right) = (\mathbf{F}\cdot\mathbf{v})\,. \tag{5.32}$$

How can we understand this result? Let the particle be displaced over a distance $\delta\mathbf{r}$ in a time increment δt. The work done by the force is given by $(\mathbf{F}\cdot\delta\mathbf{r})$. This means that the work per unit time is given by $(\mathbf{F}\cdot\delta\mathbf{r})/\delta t$. In the limit $\delta t \to 0$ the quantity $\delta\mathbf{r}/\delta t$ is the velocity $\mathbf{v}$, so that the power is given by $(\mathbf{F}\cdot\mathbf{v})$ as stated in (5.32). In Section 11.3 we will derive the time derivative of the kinetic energy of a gas or fluid, and it will be shown in (11.22) that for such a system the energy delivered by the force per unit time is also given by $(\mathbf{F}\cdot\mathbf{v})$, with $\mathbf{F}$ the force per unit volume.

5.5 Total and partial time derivatives

We have seen in the previous section that the potential that acts on a particle changes with time when the particle moves to a different location. This principle of a temporal change that is caused by a motion in the system is very general. As an example, consider the situation in Figure 5.7 where an observer measures the temperature in an observation tower. The motion of the wind is towards the left, and a region of warm air is transported leftward with the wind. The observer detects an increase of the temperature with time, so that $\partial T/\partial t > 0$. In this expression the

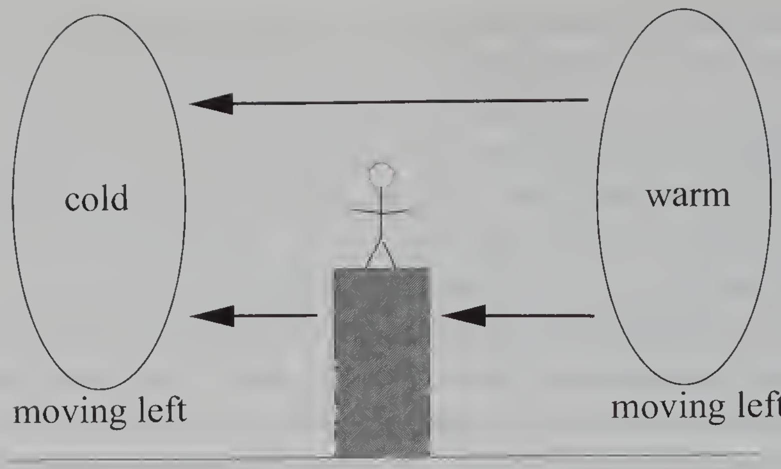

Fig. 5.7 An observer standing on a fixed tower experiences an increase in temperature because warm air moves towards the observer.

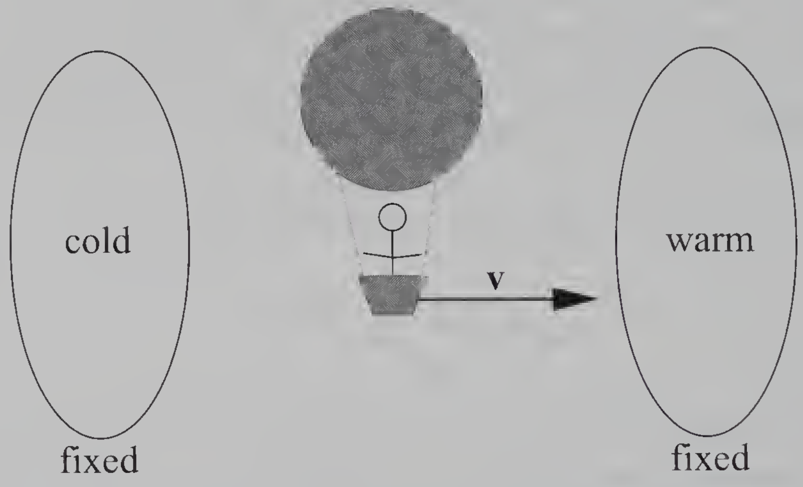

Fig. 5.8 An observer in a balloon flies through a temperature field that is constant at each location. Because the balloon flies to a warmer region the observer detects an increase in temperature.

partial derivative symbol is used deliberately. In general, the temperature field is a function of both time and the space coordinates: $T = T(\mathbf{r}, t)$. For the observer in the observation tower the space coordinates are fixed, which means that any change detected by the observer is due to the temporal change of the local temperature only, but that the location is fixed. This is expressed by the definition of the partial derivative:

$$\frac{\partial T(\mathbf{r}, t)}{\partial t} \equiv \lim_{\delta t \to 0} \frac{T(\mathbf{r}, t + \delta t) - T(\mathbf{r}, t)}{\delta t}. \tag{5.33}$$

In this derivative the time is varied but all other variables are kept fixed.

In contrast to the previous example let us now consider an observer that is carried along in a balloon by the wind, see Figure 5.8. We assume in this situation that the temperature field is fixed in time, so that $\partial T/\partial t = 0$. The observer is carried by the wind from a cold region to a warmer region. This means that just like the observer in Figure 5.7 this second observer experiences an increase in the temperature. The rate of change of the temperature is denoted by the total time derivative dT/dt. We can see from this example that dT/dt can be larger than zero while $\partial T/\partial t = 0$. The total time derivative is defined in the following way:

$$\frac{dT(\mathbf{r},t)}{dt} = \lim_{\delta t \to 0} \frac{T(\mathbf{r}(t+\delta), t+\delta t) - T(\mathbf{r}(t), t)}{\delta t}. \tag{5.34}$$

The only difference between this and the partial time derivative (5.33) is that in (5.34) the position of the observation point changes, whereas in the partial time derivative (5.33) the position is fixed.

Problem a Assuming that the temperature field in Figure 5.8 depends on position but not on time, show that dT/dt has the same form as expression (5.26) and use the results of Section 5.4 to show that for the observer in Figure 5.8

$$\frac{dT(\mathbf{r})}{dt} = (\mathbf{v} \cdot \nabla T) \qquad \text{(in the special case of Figure 5.8).} \tag{5.35}$$

Problem b Draw the gradient vector ∇T in Figure 5.8 at the location of the observer and show from the geometry of this vector and the velocity vector that $dT/dt > 0$.

This means that both observers in Figures 5.7 and 5.8 detect a rise in temperature, but their description of this temperature change is completely different. The first observer feels an increase in temperature because the local temperature changes at that observer's location; this is described by the partial time derivative $\partial T/\partial t$. The second observer detects an increase in temperature because that second observer is transported to a warmer region in a fixed temperature field; this is describes by the total time derivative that is for this special case given by $dT/dt = (\mathbf{v} \cdot \nabla T)$.

In general the temperature field may change because of a combination of a change in the temperature at a fixed location and a movement of the observer. We will assume now that the temperature field depends on all three space coordinates and on time: $T = T(x(t), y(t), z(t), t)$. In this notation it is explicit that the space coordinates of the observation point depend on time as well. The total time derivative is given by (5.34) and can be written as:

$$\frac{dT}{dt} = \lim_{\delta t \to 0} \frac{T(x(t+\delta t), y(t+\delta t), z(t+\delta t), t+\delta t) - T(x(t), y(t), z(t), t)}{\delta t}. \tag{5.36}$$

Let the x-coordinate over the time interval change from x to $x + \delta x$, and the y- and z-coordinates change in a similar way.

Problem c Use the first order Taylor series to show that

$$\delta x = x(t + \delta t) - x(t) \approx \frac{\partial x}{\partial t}\delta t = v_x \delta t. \tag{5.37}$$

In the limit $\delta t \to 0$ the approximation becomes an identity. Explain that v_x is the x-component of the velocity vector.

Problem d Generalize (3.22) for the case of a function that depends both on x, y, z, and on t to the following first order Taylor series:

$$\begin{aligned} T(x + \delta x, y + \delta y, z + \delta z, t + \delta t) &\approx T(z, y, z, t) \\ &+ \frac{\partial T}{\partial x}\delta x + \frac{\partial T}{\partial y}\delta y + \frac{\partial T}{\partial z}\delta z + \frac{\partial T}{\partial t}\delta t. \end{aligned} \tag{5.38}$$

Problem e Insert (5.37) and (5.38) in (5.36) and take the limit $\delta t \to 0$ to derive that

$$\frac{dT}{dt} = \frac{\partial T}{\partial x}v_x + \frac{\partial T}{\partial y}v_y + \frac{\partial T}{\partial z}v_z + \frac{\partial T}{\partial t}. \tag{5.39}$$

Problem f Use the definition of the gradient vector to rewrite this as

$$\frac{dT}{dt} = \frac{\partial T}{\partial t} + (\mathbf{v} \cdot \nabla T). \tag{5.40}$$

Problem g This is the general expression of the total time derivative. Show that the time derivatives (5.33) and (5.35) seen by the observers in Figures 5.7 and 5.8 can both be obtained from this general expression for the total time derivative.

The temperature field was used in this section only for illustrative purposes. The analysis is of course applicable to any function that depends on both time and the space coordinates. Also, in the analysis we used "observers" to fix our minds. However, in general there are no observers and it is not essential that there is anybody present to "observe" the change in temperature. The total time derivative is always related to the movement of a certain quantity. In Section 5.4 this was the movement of a particle that is subjected to a potential $V(\mathbf{r})$. In the description of a gas it may concern the motion of a parcel of material in the gas that is carried around by the flow.

Whenever one describes a continuous system such as the motion in the atmosphere one has a choice in the way that one sets up this description. One option is to consider every quantity at a fixed location in space, and to specify how these quantities change with time. In that case the change of the properties with time is described by the partial time derivative $\partial/\partial t$. This is called an *Eulerian* description. As an alternative one may follow a certain property while it is being carried around by the flow. In that case the change of this property with time is given by the total time derivative d/dt. This is called a *Lagrangian* description. Which description is most convenient depends on the problem. In numerical applications the Eulerian description is usually most convenient because one can work with a fixed system of space coordinates and one only needs to specify how a property changes with time in that coordinate system. On the other hand, if one wants to study, for example, the spreading of pollution by wind one aims at tracking particles, and for this the Lagrangian formulation is most convenient. For this reason it is important that one can transform the physical laws back and forth between an Eulerian and a Lagrangian formulation. We return to this issue in Chapter 11.

5.6 Gradient in spherical coordinates

The gradient vector is defined in (5.9) using a system of Cartesian coordinates. In many applications it is much more convenient to use a system of curvilinear coordinates, especially spherical coordinates and cylindrical coordinates. It is therefore useful to obtain the form of the gradient in these coordinate systems as well. In this section we consider the expression of the gradient in spherical coordinates. Before we do this we first rewrite the expression of the gradient in Cartesian coordinates.

Problem a Show that the gradient in Cartesian coordinates can also be written as:

$$\nabla f = \hat{\mathbf{x}}\frac{\partial f}{\partial x} + \hat{\mathbf{y}}\frac{\partial f}{\partial y} + \hat{\mathbf{z}}\frac{\partial f}{\partial z}. \tag{5.41}$$

Problem b Take the inner product of this expression with the unit vector in the x-direction to show that

$$\frac{\partial f}{\partial x} = (\hat{\mathbf{x}} \cdot \nabla f). \tag{5.42}$$

Expression (4.4) gives the spherical coordinates r, θ, and φ in terms of the Cartesian coordinates x, y, and z. Using the chain rule of differentiation one can then use for example that

$$\frac{\partial}{\partial x} = \frac{\partial r}{\partial x}\frac{\partial}{\partial r} + \frac{\partial \theta}{\partial x}\frac{\partial}{\partial \theta} + \frac{\partial \varphi}{\partial x}\frac{\partial}{\partial \varphi}.$$

Together with expression (4.7) for the unit vectors $\hat{\mathbf{r}}$, $\hat{\boldsymbol{\theta}}$, and $\hat{\boldsymbol{\varphi}}$ this could be used to derive the expression for the gradient in spherical coordinates. However, this approach is algebraically very complex and does not give much insight. Here we derive the gradient vector in spherical coordinates using (5.8). The idea is that the component of the gradient along a certain coordinate axis is simply given by the rate of change of the function in that direction: $\delta f = (\nabla f \cdot \delta \mathbf{s})$. Suppose that the change in position $\delta \mathbf{s}$ is in the direction of a unit vector $\hat{\mathbf{e}}$, then the change in position is given by $\delta \mathbf{s} = \hat{\mathbf{e}}\, \delta s$, where δs is given by: $\delta s \equiv |\delta \mathbf{s}|$. This means that $\delta f = (\nabla f \cdot \hat{\mathbf{e}})\, \delta s$. However, as $(\nabla f \cdot \hat{\mathbf{e}})$ is the component of the gradient vector in the direction of the unit vector $\hat{\mathbf{e}}$, this means that the component $\nabla_e f$ of the gradient vector in the direction of $\hat{\mathbf{e}}$ is given by:

$$\nabla_e f \equiv (\nabla f \cdot \hat{\mathbf{e}}) = \frac{\delta f}{\delta s} = \lim_{\delta s \to 0} \frac{f(\mathbf{r} + \hat{\mathbf{e}}\delta s) - f(\mathbf{r})}{\delta s}. \tag{5.43}$$

Consider the spherical coordinate system shown in Figure 5.9. Let us move in the direction of the unit vector $\hat{\boldsymbol{\theta}}$ from the point A with spherical coordinates (r, θ, φ) to point B with spherical coordinates $(r, \theta + \delta\theta, \varphi)$.

Problem c Derive from Figure 5.9 that the distance δs between the points A and B is related to the change in the angle $\delta\theta$ by the relation $\delta s = r\delta\theta$.

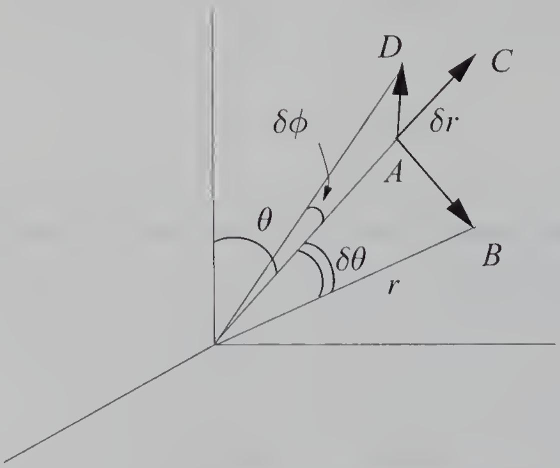

Fig. 5.9 Definition of the geometric variables and the points A, B, C, and D in spherical coordinates.

Problem d Use this in (5.43) to derive that the θ-component of the gradient vector is given by

$$\nabla_\theta f = \lim_{\delta\theta \to 0} \frac{f(r, \theta + \delta\theta, \varphi) - f(r, \theta, \varphi)}{r\delta\theta} = \frac{1}{r}\frac{\partial f}{\partial \theta}. \tag{5.44}$$

Problem e For a change from point A to point C in the $\hat{\mathbf{r}}$ direction over an increment δr in the radius the change in the position has magnitude $\delta s = \delta r$. Derive from this that the radial component is given by

$$\nabla_r f = \frac{\partial f}{\partial r}. \tag{5.45}$$

Problem f Use the geometry of Figure 5.9 to show that for a change from point A to point D in the $\hat{\boldsymbol{\varphi}}$-direction over an increment $\delta\varphi$ in the azimuth φ the change in the position has magnitude $\delta s = r\ \sin\theta\ \delta\varphi$. Derive from this that the azimuthal component of the gradient is given by

$$\nabla_\varphi f = \frac{1}{r\sin\theta}\frac{\partial f}{\partial \varphi}. \tag{5.46}$$

Combining these results we obtain the expression for the gradient in spherical coordinates:

$$\nabla f = \hat{\mathbf{r}}\frac{\partial f}{\partial r} + \hat{\boldsymbol{\theta}}\frac{1}{r}\frac{\partial f}{\partial \theta} + \hat{\boldsymbol{\varphi}}\frac{1}{r\sin\theta}\frac{\partial f}{\partial \varphi}. \tag{5.47}$$

Note that the gradient in spherical coordinates is *not* given by $\nabla f = \hat{\mathbf{r}}\partial f/\partial r + \hat{\boldsymbol{\theta}}\partial f/\partial\theta + \hat{\boldsymbol{\varphi}}\partial f/\partial\varphi$. There is a simple reason why this expression must be wrong. The first term on the right-hand side has the dimension $f/length$ because the radius is a length, while the second term has the dimension f because the angle θ is expressed in radians which is a dimensionless property. This means that this expression must be wrong. The factors $1/r$ and $1/r\sin\theta$ in (5.47) account for the scaling of the distance δs with the change in the angles $\delta\theta$ and $\delta\varphi$, respectively. These terms therefore account for the fact that the system is curvilinear.

Problem g Show that each of the terms in (5.47) has the dimension $f/length$.

Problem h Do the same analysis for cylindrical coordinates and show that the gradient in cylindrical coordinates is given by

$$\nabla f = \hat{\mathbf{r}}\frac{\partial f}{\partial r} + \hat{\boldsymbol{\varphi}}\frac{1}{r}\frac{\partial f}{\partial \varphi} + \hat{\mathbf{z}}\frac{\partial f}{\partial z}. \tag{5.48}$$

6

Divergence of a vector field

The physical meaning of the divergence cannot be understood without understanding what the flux of a vector field is, and what the sources and sinks of a vector field are.

6.1 Flux of a vector field

To fix our mind, let us consider a vector field $\mathbf{v}(\mathbf{r})$ that represents the flow of a fluid that has a constant density. We define a surface S in this fluid. Of course the surface has an orientation in space, and the unit vector perpendicular to S is denoted by $\hat{\mathbf{n}}$. Infinitesimal elements of this surface are denoted with $d\mathbf{S} \equiv \hat{\mathbf{n}} dS$. Now suppose we are interested in the volume of fluid that flows per unit time through the surface S; this quantity is called Φ. When we want to know the flow through the surface, we only need to consider the component of $\mathbf{v}$ perpendicular to the surface, the flow along the surface is not relevant.

Problem a Show that the component of the flow *across* the surface is given by $(\mathbf{v} \cdot \hat{\mathbf{n}})\hat{\mathbf{n}}$ and that the flow *along* the surface is given by $\mathbf{v} - (\mathbf{v} \cdot \hat{\mathbf{n}})\hat{\mathbf{n}}$. If you find this problem difficult you may want to look ahead to Section 13.1.

Using this result the volume of the flow through the surface per unit time is given by:

$$\Phi_{\mathrm{v}} = \iint (\mathbf{v} \cdot \hat{\mathbf{n}})\, dS = \iint \mathbf{v} \cdot d\mathbf{S}; \tag{6.1}$$

this expression defines the flux Φ_{v} of the vector field $\mathbf{v}$ through the surface S. The definition of a flux is not restricted to the flow of fluids: a flux can be computed for any vector field. However, the analogy of fluid flow often is useful for understanding the meaning of the flux and divergence.

The electric field generated by a point charge q at the origin is given by

$$\mathbf{E}(\mathbf{r}) = \frac{q\hat{\mathbf{r}}}{4\pi\varepsilon_0 r^2}. \tag{6.2}$$

In this expression $\hat{\mathbf{r}}$ is the unit vector in the radial direction and ε_0 is the permittivity. A surface element on a sphere of radius R is given by $dS = R^2 d\Omega$, with $d\Omega = \sin\theta d\theta d\varphi$ an increment in the solid angle on the unit sphere. The normal to the spherical surface is given by $\hat{\mathbf{r}}$, so that the oriented surface element is given by $d\mathbf{S} = \hat{\mathbf{r}} R^2 d\Omega$.

Problem b Show that the flux of the electric field through a spherical surface of radius R with the point charge at its center is given by

$$\Phi_{\mathrm{E}} = \iint \mathbf{E}\cdot d\mathbf{S} = \frac{q}{4\pi\varepsilon_0 R^2}\iint (\hat{\mathbf{r}}\cdot\hat{\mathbf{r}})\, R^2 d\Omega = \frac{q}{4\pi\varepsilon_0}\int_0^{2\pi}\int_0^{\pi} \sin\theta d\theta d\varphi. \tag{6.3}$$

The last identity implies that this flux is independent of the radius R.

Problem c Carry out the integration over the angles to show that the flux is given by

$$\Phi_{\mathrm{E}} = \frac{q}{\varepsilon_0}. \tag{6.4}$$

This means that the flux depends on the charge q and the permittivity ε_0, but not on the radius of the surface.

As a next example we compute the flux of the magnetic field of the Earth through the Earth's surface. The radius of the Earth is denoted by R. To first order the magnetic field of the Earth is a dipole field. (This is the field generated by a magnetic north pole and magnetic south pole that are very close together.) The dipole vector $\mathbf{m}$ points from the south pole of the dipole to the north pole and its size is given by the strength of the dipole. The magnetic field $\mathbf{B}(\mathbf{r})$ is given by (reference [53], p. 182):

$$\mathbf{B}(\mathbf{r}) = \frac{3\hat{\mathbf{r}}(\hat{\mathbf{r}}\cdot\mathbf{m}) - \mathbf{m}}{r^3}. \tag{6.5}$$

We are free to choose the orientation of the coordinate axes. Let us align the z-axis of the coordinate system by the dipole vector, in that case $\mathbf{m} = m\hat{\mathbf{z}}$, and

$$\mathbf{B}(\mathbf{r}) = m\frac{3\hat{\mathbf{r}}(\hat{\mathbf{r}}\cdot\hat{\mathbf{z}}) - \hat{\mathbf{z}}}{r^3}. \tag{6.6}$$

Problem d According to Figure 4.1, $(\hat{\mathbf{r}} \cdot \hat{\mathbf{z}}) = \cos\theta$. Use this result and the relation $d\mathbf{S} = \hat{\mathbf{r}} R^2 \sin\theta d\theta d\varphi$ to show that the magnetic flux is given by

$$\Phi_{\mathbf{B}} = m \int_0^{2\pi} \int_0^{\pi} \frac{3\cos\theta - \cos\theta}{R^3} R^2 \sin\theta d\theta d\varphi. \tag{6.7}$$

Problem e Carry out the integration over the angle θ to show that this flux vanishes.

This means that the total magnetic flux through the surface of the Earth vanishes. Physically this implies that as many magnetic field lines point out of the Earth as into the Earth. The electric field of a positive charge points away from the charge. As shown in expression (6.4) this gives a nonzero flux of the electric field through a spherical surface around that charge. This situation is different for the magnetic field where field lines point away from the dipole as well as towards the dipole, which causes the magnetic flux to vanish.

6.2 Introduction of the divergence

In order to introduce the divergence, consider an infinitesimal rectangular volume with sides dx, dy, and dz (see Figure 6.1 for the definition of the geometric variables). The outward flux through the right surface perpendicular through the x-axis is given by $v_x(x + dx, y, z)dydz$, because $v_x(x + dx, y, z)$ is the component of the flow perpendicular to that surface and $dydz$ is the area of the surface. By the same token, the flux through the left surface perpendicular through the x-axis is given by

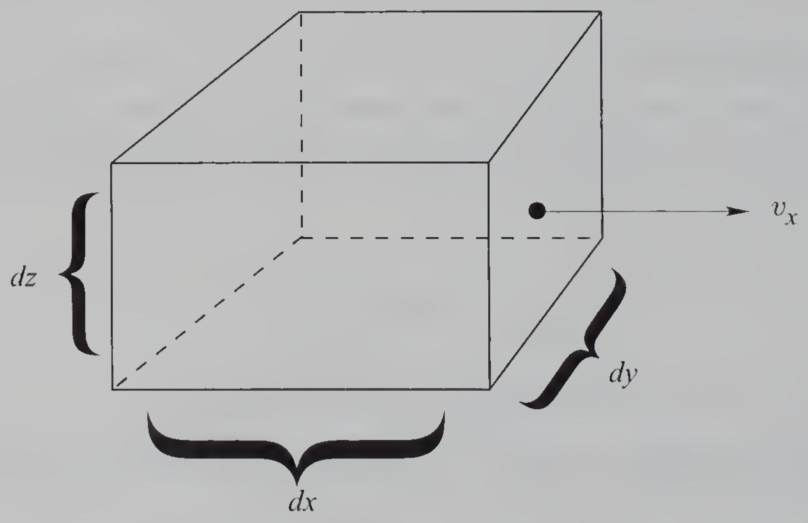

Fig. 6.1 Definition of the geometric variables in the calculation of the flux of a vector field through an infinitesimal rectangular volume.

$-v_x(x, y, z)dydz$, the minus sign is due to the fact that the component of $\mathbf{v}$ in the direction *outward* from the cube is given by $-v_x$. (Alternatively one can say that for this surface the unit vector perpendicular to the surface and pointing outwards is given by $\hat{\mathbf{n}} = -\hat{\mathbf{x}}$.) This means that the total outward flux through the two surfaces is given by $v_x(x + dx, y, z)dydz - v_x(x, y, z)dydz = (\partial v_x/\partial x)\, dxdydz$. The same reasoning applies to the surfaces perpendicular to the y- and z-axes. This means that the total outward flux through the sides of the cubes is:

$$d\Phi_{\mathbf{v}} = \left(\frac{\partial v_x}{\partial x} + \frac{\partial v_y}{\partial y} + \frac{\partial v_z}{\partial z}\right) dV = (\nabla \cdot \mathbf{v})\, dV, \tag{6.8}$$

where dV is the volume $dxdydz$ of the cube and $(\nabla \cdot \mathbf{v})$ is the *divergence* of the vector field $\mathbf{v}$. The above definition does not really tell us yet what the divergence really is. Dividing (6.8) by dV one obtains $(\nabla \cdot \mathbf{v}) = d\Phi_{\mathbf{v}}/dV$. This allows us to state in words what the divergence is:

> *The divergence of a vector field is the outward flux of the vector field per unit volume.*

To fix our minds again let us consider a physical example in two dimensions in which a fluid is pumped into this two-dimensional space at the origin $\mathbf{r} = 0$. For simplicity we assume that the fluid is incompressible, which means that the mass-density is constant. We do not know yet what the resulting flow field is, but we know two things. Away from the source at $\mathbf{r} = 0$ there are no sources or sinks of fluid flow. This means that the flux of the flow through any closed surface S that does not enclose any sources or sinks must be zero: "what goes in must come out." This in turn means that the divergence of the flow is zero, except possibly near the source at $\mathbf{r} = 0$:

$$(\nabla \cdot \mathbf{v}) = 0 \qquad \text{for} \quad \mathbf{r} \neq 0. \tag{6.9}$$

In addition we know that due to the symmetry of the problem the flow is directed in the radial direction and depends on the radius r only:

$$\mathbf{v}(\mathbf{r}) = f(r)\mathbf{r}. \tag{6.10}$$

This is enough information to determine the flow field. Of course, it is a problem that we cannot immediately insert (6.10) in (6.9) because we have not yet derived an expression for the divergence in cylindrical coordinates. However, there is another way to determine the flow from the expression above.

Problem a Use the identity $r = \sqrt{x^2 + y^2}$ to show that

$$\frac{\partial r}{\partial x} = \frac{x}{r}, \tag{6.11}$$

and derive the corresponding equation for y. Using expressions (6.10), (6.11), and the chain rule for differentiation show that

$$(\nabla \cdot \mathbf{v}) = 2f(r) + r\frac{df}{dr} \qquad \text{(cylindrical coordinates)}\,. \tag{6.12}$$

Problem b Insert this result in (6.9) and show that the flow field is given by $\mathbf{v}(\mathbf{r}) = A\mathbf{r}/r^2$. Make a sketch of the flow field.

The constant A is yet to be determined. Let a volume of liquid $\dot{V}$ be injected per unit time at the source $\mathbf{r} = 0$. (Following an often used convention, the dot denotes a time derivative.)

Problem c Show that $\dot{V} = \int \mathbf{v} \cdot d\mathbf{S}$ (where the integration is over an arbitrary surface around the source at $\mathbf{r} = 0$). Choosing a suitable surface, derive that

$$\mathbf{v}(\mathbf{r}) = \frac{\dot{V}}{2\pi}\frac{\hat{\mathbf{r}}}{r}. \tag{6.13}$$

Note that the unit vector $\hat{\mathbf{r}}$ is now used rather than the position vector $\mathbf{r}$. These vectors are related by the expression $\mathbf{r} = r\hat{\mathbf{r}}$.

From this simple example of a single source at $\mathbf{r} = 0$ more complex examples can be obtained. Suppose we have a source at $\mathbf{r}_+ = (L, 0)$, where a volume $\dot{V}$ is injected per unit time, and a sink at $\mathbf{r}_- = (-L, 0)$, where a volume $-\dot{V}$ is removed per unit time. The total flow field can be obtained by superposition of flow fields of the form (6.13) for the source and the sink. The flow field from the source at $\mathbf{r}_+$ can be found by replacing the location vector $\mathbf{r}$ in (6.13) by the vector $\mathbf{r} - \mathbf{r}_+$. This amounts to the following replacements:

$$r = \sqrt{x^2 + y^2} \rightarrow \sqrt{(x - L)^2 + y^2}, \tag{6.14}$$

and

$$\hat{\mathbf{r}} = \mathbf{r}/r = \begin{pmatrix} x \\ y \end{pmatrix} \Big/ \sqrt{x^2 + y^2} \rightarrow \begin{pmatrix} x - L \\ y \end{pmatrix} \Big/ \sqrt{(x - L)^2 + y^2}. \tag{6.15}$$

Problem d Make similar replacements for the sink at $\mathbf{r}_-$ and show that the x- and y-components of the flow field in this case are given by:

$$v_x(x, y) = \frac{\dot{V}}{2\pi}\left[\frac{x - L}{(x - L)^2 + y^2} - \frac{x + L}{(x + L)^2 + y^2}\right], \tag{6.16}$$

$$v_y(x, y) = \frac{\dot{V}}{2\pi}\left[\frac{y}{(x - L)^2 + y^2} - \frac{y}{(x + L)^2 + y^2}\right], \tag{6.17}$$

and sketch the resulting flow field. This is most easily accomplished by determining from these expressions the flow field at some selected lines such as the x- and y-axes.

One may also be interested in computing the *streamlines* of the flow. These are the lines along which material particles flow. The streamlines can be found by using that the time derivative of the position of a material particle is the velocity:

$$d\mathbf{r}/dt = \mathbf{v}(\mathbf{r}). \tag{6.18}$$

Inserting expressions (6.16) and (6.17) into equation (6.18) leads to two coupled differential equations for $x(t)$ and $y(t)$ which are difficult to solve. Fortunately, there are more intelligent ways of retrieving the streamlines. We return to this issue in Section 16.3.

6.3 Sources and sinks

In the above example, the fluid flow moves away from the source and converges on the sink. The terms "source" and "sink" have a clear physical meaning since they are directly related to the "source" of water as from a tap, and a "sink" as the sink in a bathtub. The flow lines of the water flow diverge from the source while they converge towards the sink. This explains the term "divergence", because this quantity simply indicates to what extent flow lines originate (in the case of a source) or end (in the case of a sink).

This definition of sources and sinks is not restricted to fluid flow. For example, for the electric field the term "fluid flow" should be replaced by the term "field lines." Electrical field lines originate at positive charges and end at negative charges. We first verify this by computing the divergence of the electrical field (6.2) for a point charge in three dimensions.

Problem a Show that expression (6.2) can be written as

$$\mathbf{E}(\mathbf{r}) = \frac{q}{4\pi\varepsilon_0 r^3}\begin{pmatrix} x \\ y \\ z \end{pmatrix}. \tag{6.19}$$

Problem b The x-derivative of the x-component of the electric field gives a contribution

$$\frac{\partial}{\partial x}\frac{x}{r^3} = \frac{1}{r^3} - \frac{3x}{r^4}\frac{\partial r}{\partial x}. \tag{6.20}$$

Use expression (6.11) to show that this quantity is given by

$$\frac{\partial}{\partial x}\frac{x}{r^3} = \frac{1}{r^5}\left(r^2 - 3x^2\right). \tag{6.21}$$

Problem c Compute the y, and z-derivatives in the divergence of the electric field as well and show that

$$(\nabla \cdot \mathbf{E}) = \frac{q}{4\pi\varepsilon_0 r^5}\left(3r^3 - 3(x^2 + y^2 + z^2)\right) = 0. \tag{6.22}$$

This means that away from the point charge ($r \neq 0$) the divergence of the electric charge vanishes. At the charge, $r = 0$ and it is not clear what the right-hand side of expression (6.22) is. It follows from equation (6.4) that the net flux through a spherical surface is given by q/ε_0. This flux is nonzero when the charge q is nonzero. For a positive charge the flux is positive, while for a negative charge the flux is negative. According to expression (6.22) the divergence vanishes for $r \neq 0$. The flux can be nonzero only when the divergence is nonzero at the location of the point charge.

Physically this means that the electric charge is the source of the electric field. This is reflected in the Maxwell equation for the divergence of the electric field:

$$(\nabla \cdot \mathbf{E}) = \rho(\mathbf{r})/\varepsilon_0. \tag{6.23}$$

In this expression $\rho(\mathbf{r})$ is the charge density, which is simply the electric charge per unit volume just as the mass-density denotes the mass per unit volume. In addition, expression (6.23) contains the permittivity ε_0. This term serves as a *coupling constant* since it describes how "much" electrical field is generated by a given electrical charge density. It is obvious that a constant is needed here because the charge density and the electrical field have different physical dimensions, hence a proportionality factor must be present. However, the physical meaning of a coupling constant goes much deeper, because it prescribes how strong the field is that is generated by a given source. This constant describes how strongly cause (the source) and effect (the field) are coupled.

As the next example we compute the divergence of the magnetic field (6.5) generated by a dipole $\mathbf{m}$. We are free to choose the orientation of the coordinate system. In the following we choose the z-axis to be aligned with the dipole vector $\mathbf{m}$, this means that $\mathbf{m} = m\hat{\mathbf{z}}$.

Problem d Use this expression and the relation $\hat{\mathbf{r}} = \mathbf{r}/r$ to show that the magnetic field (6.5) can be written as

$$\mathbf{B} = \frac{3m\mathbf{r}\,(\mathbf{r}\cdot\hat{\mathbf{z}})}{r^5} - \frac{m\hat{\mathbf{z}}}{r^3}. \tag{6.24}$$

Problem e Write the vectors $\mathbf{r}$ and $\hat{\mathbf{z}}$ in component form to show that the components of the magnetic field are given by

$$\mathbf{B} = \begin{pmatrix} 3mxz/r^5 \\ 3myz/r^5 \\ 3mz^2/r^5 - m/r^3 \end{pmatrix} . \tag{6.25}$$

Problem f Differentiate the different components to show that

$$\begin{aligned} \frac{\partial B_x}{\partial x} &= 3m\left(\frac{z}{r^5} - \frac{5x^2 z}{r^7}\right), \\ \frac{\partial B_y}{\partial y} &= 3m\left(\frac{z}{r^5} - \frac{5y^2 z}{r^7}\right), \\ \frac{\partial B_z}{\partial z} &= 3m\left(\frac{3z}{r^5} - \frac{5z^3}{r^7}\right). \end{aligned} \tag{6.26}$$

Problem g Add these terms to show that the divergence of the magnetic field vanishes:

$$(\nabla \cdot \mathbf{B}) = 0. \tag{6.27}$$

By analogy with (6.23) one might expect that the divergence of the magnetic field is nonzero at the source of the field and that it is related to a magnetic charge density:

$$(\nabla \cdot \mathbf{B}) = \textit{coupling const.} \times \rho_B(\mathbf{r}),$$

where ρ_B would be the "density of magnetic charge." However, particles with a magnetic charge (usually called "magnetic monopoles") have not been found in nature despite extensive searches. Therefore the Maxwell equation for the divergence of the magnetic field is:

$$(\nabla \cdot \mathbf{B}) = 0, \tag{6.28}$$

but we should remember that this divergence is zero because of the observational absence of magnetic monopoles rather than a vanishing coupling constant.

6.4 Divergence in cylindrical coordinates

In the previous analysis we have expressed the divergence in Cartesian coordinates: $\nabla \cdot \mathbf{v} = \partial_x v_x + \partial_y v_y + \partial_z v_z$. As you may have discovered, the use of other coordinate systems such as cylindrical coordinates or spherical coordinates can often make

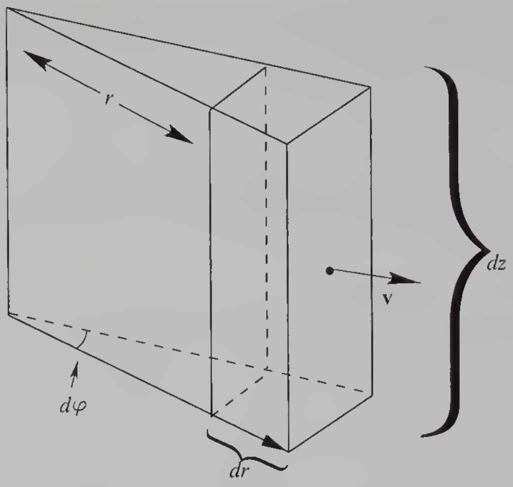

Fig. 6.2 Definition of the geometric variables for the computation of the divergence in cylindrical coordinates.

life much simpler. Here we derive an expression for the divergence in cylindrical coordinates. In this system, the distance $r = \sqrt{x^2 + y^2}$ of a point to the z-axis, the azimuth $\varphi (= \arctan(y/x))$ and z are used as coordinates, see Section 4.5. A vector $\mathbf{v}$ can be decomposed into components in this coordinate system:

$$\mathbf{v} = v_r \hat{\mathbf{r}} + v_\varphi \hat{\boldsymbol{\varphi}} + v_z \hat{\mathbf{z}}, \tag{6.29}$$

where $\hat{\mathbf{r}}$, $\hat{\boldsymbol{\varphi}}$, and $\hat{\mathbf{z}}$ are unit vectors in the direction of increasing values of r, φ, and z, respectively. As shown in Section 6.2 the divergence is the flux per unit volume. Let us consider the infinitesimal volume corresponding to increments dr, $d\varphi$, and dz shown in Figure 6.2. Let us first consider the flux of $\mathbf{v}$ through the surface elements perpendicular to $\hat{\mathbf{r}}$. The size of this surface is $r d\varphi dz$ and $(r + dr) d\varphi dz$ at r and $r + dr$ respectively. The normal components of $\mathbf{v}$ through these surfaces are $v_r(r, \varphi, z)$ and $v_r(r + dr, \varphi, z)$ respectively. Hence the total flux through these two surface is given by $v_r(r + dr, \varphi, z)(r + dr) d\varphi dz - v_r(r, \varphi, z)(r) d\varphi dz$.

Problem a Show that to first order in dr this quantity is equal to

$$\frac{\partial}{\partial r}(r v_r)\, dr d\varphi dz.$$

Hint: use a first order Taylor expansion for $v_r(r + dr, \varphi, z)$ in the quantity dr.

Problem b Show that the flux through the surfaces perpendicular to $\hat{\boldsymbol{\varphi}}$ is to first order in $d\varphi$ given by $(\partial v_\varphi / \partial \varphi) dr d\varphi dz$.

Problem c Show that the flux through the surfaces perpendicular to $\hat{\mathbf{z}}$ is to first order in dz given by $(\partial v_z / \partial z) r dr d\varphi dz$.

Problem d Use Figure 6.2 to show that the volume dV is given by

$$dV = rdrd\varphi dz. \tag{6.30}$$

Problem e Use the fact that the divergence is the flux per unit volume, and the results of the previous problems to show that in cylindrical coordinates:

$$(\nabla \cdot \mathbf{v}) = \frac{1}{r}\frac{\partial}{\partial r}(rv_r) + \frac{1}{r}\frac{\partial v_\varphi}{\partial \varphi} + \frac{\partial v_z}{\partial z}. \tag{6.31}$$

Problem f Use this result to re-derive (6.12) without using Cartesian coordinates as an intermediary. Make sure not to confuse $\mathbf{r}$ with $\hat{\mathbf{r}}$.

In spherical coordinates a vector $\mathbf{v}$ can be expanded into the components v_r, v_θ, and v_φ in the directions of increasing values of r, θ, and φ respectively. In this coordinate system r has a different meaning to that in cylindrical coordinates because in spherical coordinates $r = \sqrt{x^2 + y^2 + z^2}$.

Problem g Use the same reasoning as for the cylindrical coordinates to show that in spherical coordinates

$$\nabla \cdot \mathbf{v} = \frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2 v_r\right) + \frac{1}{r\sin\theta}\frac{\partial}{\partial \theta}(\sin\theta\ v_\theta) + \frac{1}{r\sin\theta}\frac{\partial v_\varphi}{\partial \varphi}. \tag{6.32}$$

6.5 Is life possible in a five-dimensional world?

In this section we investigate whether the motion of the Earth around the Sun is stable or not. This means that we ask ourselves whether, when the position of the Earth is perturbed (for example by the gravitational attraction of the other planets or by a passing asteroid), the gravitational force causes the Earth to return to its original position (stability) or to spiral away from (or towards) the Sun. It turns out that these stability properties depend on the spatial dimension! We know that we live in a world of three spatial dimensions, but it is interesting to investigate whether the orbit of the Earth would also be stable in a world with a different number of spatial dimensions.

In the Newtonian theory the gravitational field $\mathbf{g}(\mathbf{r})$ satisfies (see reference [76]):

$$(\nabla \cdot \mathbf{g}) = -4\pi G\rho, \tag{6.33}$$

where $\rho(\mathbf{r})$ is the mass-density and G is the gravitational constant which has a value of 6.67×10^{-8} cm^3 g^{-1} s^{-2}. The term G plays the role of a coupling constant, just like the 1/permittivity in (6.23). Note that the right-hand side of the gravitational field equation (6.33) has the opposite sign to the right-hand side of the electric field equation (6.23). This is due to the fact that two electric charges of equal sign repel each other, while two masses of equal sign (mass being positive) attract each other. Zee [125] gives an insightful explanation based on quantum field theory of this fundamental difference between the electrical field and the gravitational field. If the sign of the right-hand side of (6.33) was positive, masses would repel each other and structures such as planets, the solar system, and stellar systems would not exist.

Problem a We argued in Section 6.3 that electric field lines start at positive charges and end at negative charges. By analogy we expect that gravitational field lines end at the (positive) masses that generate the field. However, where do the gravitational field lines start?

Let us first determine the gravitational field of the Sun in N dimensions. Outside the Sun the mass-density vanishes; this means that $(\nabla \cdot \mathbf{g}) = 0$. We assume that the mass-density in the Sun is spherically symmetric, then the gravitational field must also be spherically symmetric and is thus of the form:

$$\mathbf{g}(\mathbf{r}) = f(r)\mathbf{r}. \tag{6.34}$$

In order to make further progress we must derive the divergence of a spherically symmetric vector field in N dimensions. Generalizing expression (6.32) to an arbitrary number of dimensions is not trivial, but fortunately this is not needed. We will make use of the property that in N dimensions: $r = \sqrt{\sum_{i=1}^{N} x_i^2}$.

Problem b Derive from this expression that

$$\partial r / \partial x_j = x_j / r. \tag{6.35}$$

Use this result to derive that for a vector field of the form (6.34):

$$(\nabla \cdot \mathbf{g}) = N f(r) + r \frac{\partial f}{\partial r}. \tag{6.36}$$

Outside the Sun, where the mass-density vanishes and $(\nabla \cdot \mathbf{g}) = 0$ we can use this result to determine the gravitational field.

Problem c Derive that

$$\mathbf{g}(\mathbf{r}) = -\frac{A}{r^{N-1}}\hat{\mathbf{r}}, \tag{6.37}$$

and check this result for three spatial dimensions.

At this point the constant A is not determined, but this is not important for the coming arguments. The minus sign has been added for convenience: the gravitational field points towards the Sun and hence $A > 0$.

Associated with the gravitational field is a gravitational force that attracts the Earth towards the Sun. If the mass of the Earth is denoted by m, this force is given by

$$\mathbf{F}_{grav} = -\frac{Am}{r^{N-1}}\hat{\mathbf{r}}, \tag{6.38}$$

and is directed towards the Sun. For simplicity we assume that the Earth is in a circular orbit. This means that the attractive gravitational force is balanced by the repulsive centrifugal force which is given by

$$\mathbf{F}_{cent} = \frac{mv^2}{r}\hat{\mathbf{r}}. \tag{6.39}$$

In equilibrium these forces balance: $\mathbf{F}_{grav} + \mathbf{F}_{cent} = 0$.

Problem d Derive the velocity v from this requirement.

We now assume that the distance to the Sun changes from r to $r + \delta r$; the perturbation in the position is therefore $\delta\mathbf{r} = \delta r\ \hat{\mathbf{r}}$. Because of this perturbation, the gravitational force and the centrifugal force are also perturbed. These quantities will be denoted by $\delta\mathbf{F}_{grav}$ and $\delta\mathbf{F}_{cent}$ respectively, see Figure 6.3. The change in the total force is given by $\delta\mathbf{F}_{grav} + \delta\mathbf{F}_{cent}$. Let us first suppose that the Earth moves away from the Sun, in that case $\delta\mathbf{r}$ point from the Sun to the Earth. The orbit is stable when the net perturbation of the force draws the Earth back towards the Sun. In that case the vectors $\delta\mathbf{r}$ and $\left(\delta\mathbf{F}_{grav} + \delta\mathbf{F}_{cent}\right)$ point in opposite directions and

$$\left(\delta\mathbf{F}_{grav} + \delta\mathbf{F}_{cent}\right) \cdot \delta\mathbf{r} < 0 \qquad \text{(stability)}. \tag{6.40}$$

Suppose on the other hand that the Earth moves towards the Sun, in that case $\delta\mathbf{r}$ points towards the Sun. The orbit is stable when the perturbation of the force pushes the Earth away from Sun. In that case the vectors $\delta\mathbf{r}$ and $\left(\delta\mathbf{F}_{grav} + \delta\mathbf{F}_{cent}\right)$ point in opposite directions as well so that (6.40) is satisfied. This means that the orbital motion is stable for perturbations when the gravitational field satisfies the criterion (6.40).

In order to compute the change in the centrifugal force we use that angular momentum is conserved, that is

$$mrv = m(r + \delta r)(v + \delta v). \tag{6.41}$$

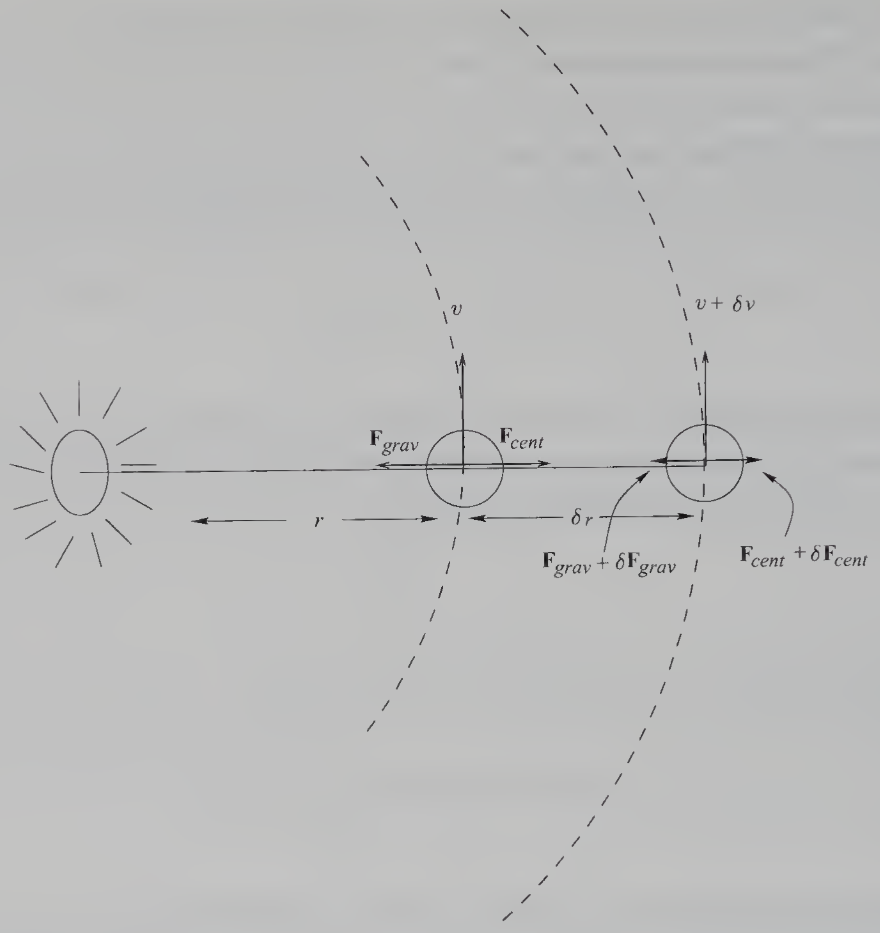
v
$v+\delta v$
$\mathbf{F}_{grav}$
$\mathbf{F}_{cent}$
r
δr
$\mathbf{F}_{grav}+\delta\mathbf{F}_{grav}$
$\mathbf{F}_{cent}+\delta\mathbf{F}_{cent}$

Fig. 6.3 Definition of variables for the perturbed orbit of the Earth.

In what follows we consider small perturbations and retain only terms of first order in the perturbation. This means that we ignore higher order terms such as the product $\delta r \delta v$.

Problem e Use equation (6.41) to derive that to first order in δr

$$\delta v = -\frac{v}{r}\delta r. \tag{6.42}$$

Problem f Use this result and expression (6.39) to derive that

$$\delta \mathbf{F}_{cent} = -\frac{3mv^2}{r^2}\delta \mathbf{r}, \tag{6.43}$$

then use (6.38) to show that

$$\delta \mathbf{F}_{grav} = (N-1)\,\frac{Am}{r^N}\delta \mathbf{r}\,. \tag{6.44}$$

Note that the perturbation of the centrifugal force does not depend on the number of spatial dimensions, but that the perturbation of the gravitational force does depend on N.

Problem g Use the value of the velocity derived in problem d and expressions (6.43)–(6.44) to show that according to the criterion (6.40) the orbital motion is stable in less than four spatial dimensions. Show also that the requirement for stability is independent of the original distance r.

This is an intriguing result. It implies that orbital motion is unstable in more than four spatial dimensions. This means that in a world with five spatial dimensions the solar system would not be stable. Life seems to be tied to planetary systems with a central star which supplies the energy to sustain life on the orbiting planet(s). This implies that life would be impossible in a five-dimensional world! Note also that the stability requirement is independent of r, that is the stability properties of orbital motion do not depend on the size of the orbit. This implies that the gravitational field does not have “stable regions” and “unstable regions”, the stability property depends only on the number of spatial dimensions.

7

Curl of a vector field

7.1 Introduction of the curl

We introduce the *curl* of a vector field $\mathbf{v}$ by its formal definition in terms of Cartesian coordinates (x, y, z) and unit vectors $\hat{\mathbf{x}}$, $\hat{\mathbf{y}}$, and $\hat{\mathbf{z}}$ in the x-, y-, and z-directions respectively:

$$curl\ \mathbf{v} \equiv \begin{vmatrix} \hat{\mathbf{x}} & \hat{\mathbf{y}} & \hat{\mathbf{z}} \\ \partial_x & \partial_y & \partial_z \\ v_x & v_y & v_z \end{vmatrix} = \begin{pmatrix} \partial_y v_z - \partial_z v_y \\ \partial_z v_x - \partial_x v_z \\ \partial_x v_y - \partial_y v_x \end{pmatrix} . \tag{7.1}$$

It can be seen that the curl is a *vector*, in contrast to the divergence which is a scalar. The notation with the determinant is sloppy because the entries in a determinant should be numbers rather than vectors such as $\hat{\mathbf{x}}$ or differentiation operators such as $\partial_y = \partial/\partial y$. However, the notation in terms of a determinant is a simple way to remember the definition of the curl in Cartesian coordinates. We will write the curl of a vector field also as: curl $\mathbf{v} = \nabla \times \mathbf{v}$.

Problem a Verify that this notation with the curl expressed as the outer product of the operator ∇ and the vector $\mathbf{v}$ is consistent with definition (7.1).

In general the curl is a three-dimensional vector. To see the physical interpretation of the curl, we will make life easy for ourselves by choosing a Cartesian coordinate system in which the z-axis is aligned with curl $\mathbf{v}$. In that coordinate system the curl is given by: curl $\mathbf{v} = (\partial_x v_y - \partial_y v_x)\hat{\mathbf{z}}$. Consider a little rectangular surface element oriented perpendicular to the z-axis with sides dx and dy respectively, see Figure 7.1. We consider the line integral $\oint_{dxdy} \mathbf{v} \cdot d\mathbf{r}$ along a closed loop defined by the sides of this surface element, integrating in the counterclockwise direction. This line integral can be written as the sum of the integral over the four sides of the surface element.

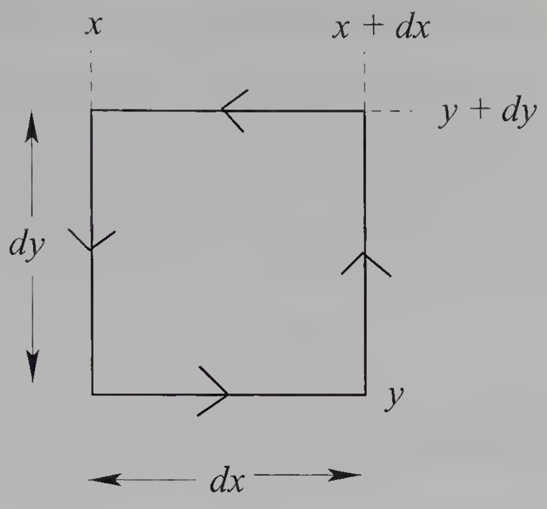

Fig. 7.1 Definition of the geometric variables for the interpretation of the curl.

Problem b Show that the line integral is given by $\oint_{dxdy} \mathbf{v} \cdot d\mathbf{r} = v_x(x, y)dx + v_y(x + dx, y)dy - v_x(x, y + dy)dx - v_y(x, y)dy$, and use a first order Taylor expansion to write this as

$$\oint_{dxdy} \mathbf{v} \cdot d\mathbf{r} = (\partial_x v_y - \partial_y v_x)\, dxdy, \tag{7.2}$$

with v_x and v_y evaluated at the point (x, y).

This expression can be rewritten as:

$$(\text{curl } \mathbf{v})_z = (\partial_x v_y - \partial_y v_x) = \frac{\oint_{dxdy} \mathbf{v} \cdot d\mathbf{r}}{dxdy}. \tag{7.3}$$

In this form we can express the meaning of the *curl* in words:

> *The component of* curl **v** *in a certain direction is the closed line integral of* **v** *along a closed path perpendicular to this direction, normalized per unit surface area.*

Note that this interpretation is similar to the interpretation of the divergence given in Section 6.2. There is, however, one major difference. The *curl* is a vector while the divergence is a scalar. This is reflected in our interpretation of the curl because a surface has an orientation defined by its normal vector, hence the curl is also a vector.

7.2 What is the curl of the vector field?

In order to discover the meaning of the curl, we consider again an incompressible fluid and study the curl of the velocity vector $\mathbf{v}$, because this will allow us to discover the physical meaning of the curl. It is not only for a didactic purpose that we consider the curl of fluid flow. In fluid mechanics this quantity plays such a crucial role that it is given a special name, the *vorticity* $\boldsymbol{\omega}$:

$$\boldsymbol{\omega} \equiv \nabla \times \mathbf{v}. \tag{7.4}$$

To simplify things further we assume that the fluid moves in the (x, y)-plane only (i.e. $v_z = 0$) and that the flow depends only on x and y: $\mathbf{v} = \mathbf{v}(x, y)$.

Problem a Show that for such a flow

$$\boldsymbol{\omega} = \nabla \times \mathbf{v} = (\partial_x v_y - \partial_y v_x)\hat{\mathbf{z}}. \tag{7.5}$$

We first consider an axisymmetric flow field. Such a flow field has rotational symmetry around one axis, and we will take this to be the z-axis. Because of the cylindrical symmetry and the fact that it is assumed that the flow does not depend on z, the components v_r, v_φ and v_z depend neither on the azimuth φ (= arctan y/x) used in the cylinder coordinates nor on z but only on the distance $r = \sqrt{x^2 + y^2}$ to the z-axis.

Problem b Show that it follows from expression (6.31) for the divergence in cylindrical coordinates that for an axisymmetric flow field for an incompressible fluid (where $(\nabla \cdot \mathbf{v}) = 0$ everywhere including the z-axis where $r = 0$) the radial component of the velocity must vanish: $v_r = 0$. Hint: for such a flow field the φ-derivative and the z-derivative both vanish.

This result simply reflects that for an incompressible flow with cylindrical symmetry there can be no net flow towards (or away from) the symmetry axis. The only nonzero component of the flow is therefore in the direction of $\hat{\boldsymbol{\varphi}}$. This implies that the velocity field must be of the form:

$$\mathbf{v} = \hat{\boldsymbol{\varphi}} v(r); \tag{7.6}$$

see Figure 7.2 for a sketch of this flow field. The velocity depends only on r because the assumed symmetry implies that $\partial v/\partial\varphi = \partial v/\partial z = 0$. The problem we now face is that definition (7.1) is expressed in Cartesian coordinates while the velocity in equation (7.6) is expressed in cylindrical coordinates. In Section 7.6 an expression

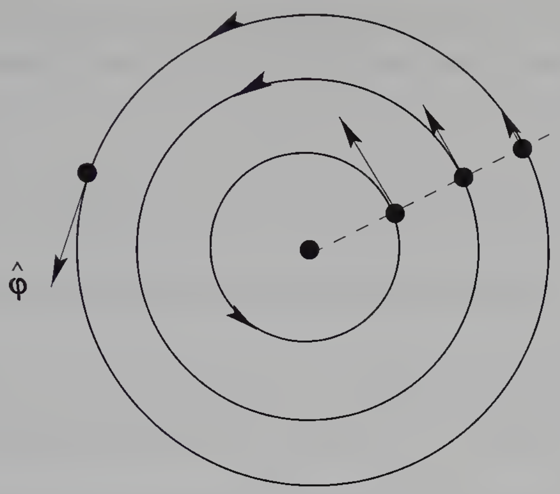

Fig. 7.2 Sketch of an axisymmetric source-free flow in the (x, y)-plane.

for the curl in cylindrical coordinates will be derived. As an alternative, one can express the unit vector $\hat{\varphi}$ in Cartesian coordinates.

Problem c Verify that:

$$\hat{\varphi} = \begin{pmatrix} -y/r \\ x/r \\ 0 \end{pmatrix}. \tag{7.7}$$

Hints: draw this vector in the (x, y)-plane, verify that this vector is perpendicular to the position vector **r** and that it is of unit length. Alternatively you can use expression (4.38).

Problem d Use (7.5), (7.7), and the chain rule for differentiation to show that for the flow field (7.6):

$$(\nabla \times \mathbf{v})_z = \frac{\partial v}{\partial r} + \frac{v}{r}. \tag{7.8}$$

Hint: you have to use the derivatives $\partial r/\partial x$ and $\partial r/\partial y$ again. You determined these in Section 6.2.

7.3 First source of vorticity: rigid rotation

In general, a nonzero curl of a vector field can have two origins: rigid rotation and shear. In this section we treat the effect of rigid rotation. Because we will use fluid flow as an example we will speak about the vorticity, but keep in mind that the results of this section (and the next) apply to any vector field. We consider a velocity field that describes a rigid rotation with the z-axis as the axis of rotation and an angular velocity Ω. This velocity field is of the form (7.6) with $v(r) = \Omega r$, because for a rigid rotation the velocity increases linearly with the distance r to the rotation axis.

Problem a Verify explicitly that every particle in the flow makes one revolution in a time $T = 2\pi/\Omega$ and that this time does not depend on the position of the particle. Hint: how long does it take to travel along the perimeter of a circle with radius r?

Problem b Use expression (7.8) to show that for this velocity field: $\nabla \times \mathbf{v} = 2\Omega\hat{\mathbf{z}}$.

This means that the vorticity is twice the rotational vector $\Omega\hat{\mathbf{z}}$. This result is derived here for the special case that the z-axis is the axis of rotation. (This can always be achieved because any orientation of the coordinate system can be chosen.) In Section 6.11 of Boas [19] it is shown with a different derivation that the vorticity for rigid rotation is given by $\boldsymbol{\omega} = \nabla \times \mathbf{v} = 2\boldsymbol{\Omega}$, where $\boldsymbol{\Omega}$ is the rotation vector. (Beware, the notation used by Boas [19] is different from that used in this book in a deceptive way!)

We see that rigid rotation leads to a vorticity that is twice the rotation rate. Imagine we place a paddle-wheel in the flow field that is associated with the rigid rotation, see Figure 7.3. This paddle-wheel moves with the flow and makes one revolution about its axis in a time $2\pi/\Omega$. Note also that for the sense of rotation shown in Figure 7.3 the paddle-wheel moves in the counterclockwise direction and that the curl points along the positive z-axis. This implies that the rotation of the paddle-wheel not only denotes that the curl is nonzero, but also that the rotation vector of the paddle is directed along the curl! This actually explains the

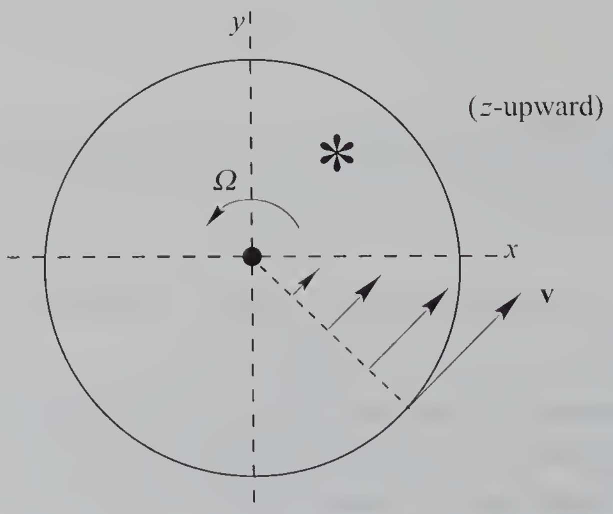

Fig. 7.3 The vorticity for a rigid rotation.

origin of the word *vorticity*. In a vortex, the flow rotates around a rotational axis. The curl increases with the rotation rate, hence it increases with the strength of the vortex. This strength of the vortex has been dubbed *vorticity,* and this term therefore reflects the fact that the curl of velocity denotes the (local) intensity of rotation in the flow.

7.4 Second source of vorticity: shear

In addition to rigid rotation, shear is another cause of vorticity. In order to see this we consider a fluid in which the flow is only in the x-direction and where the flow depends on the y-coordinate only: $v_y = v_z = 0$, $v_x = f(y)$.

Problem a Show that this flow does not describe a rigid rotation. Hint: how long does it take before a fluid particle returns to its original position?

Problem b Show that for this flow

$$\nabla \times \mathbf{v} = -\frac{\partial f}{\partial y}\hat{\mathbf{z}}. \tag{7.9}$$

As a special example consider the velocity given by:

$$v_x = f(y) = v_0 e^{-y^2/L^2}. \tag{7.10}$$

This flow field is sketched in Figure 7.4. The top paddle-wheel rotates in the counterclockwise direction because the flow at the top of the wheel is slower than the flow at the bottom of the paddle-wheel. Therefore, the paddle-wheels placed in the flow rotate in the sense indicated in Figure 7.4.

Problem c Compute $\nabla \times \mathbf{v}$ for this flow field and verify that both the curl and the rotational vector of the paddle wheels are aligned with the z-axis. Show that the vorticity is positive where the paddle-wheels rotate in the counterclockwise direction and that the vorticity is negative where the paddle-wheels rotate in the clockwise direction.

It follows from the example in this section and the example in Section 7.3 that both rotation and shear cause a nonzero vorticity. Both phenomena lead to the rotation of imaginary paddle-wheels embedded in the vector field. Therefore, the curl of a vector field measures the local rotation of the vector field (in a literal sense). This explains why in some languages (i.e. Dutch) the notation *rot* **v** is used rather than *curl* **v**. Note that this interpretation of the *curl* as a measure of (local) rotation is consistent with (7.3) in which the curl is related to the value of the line

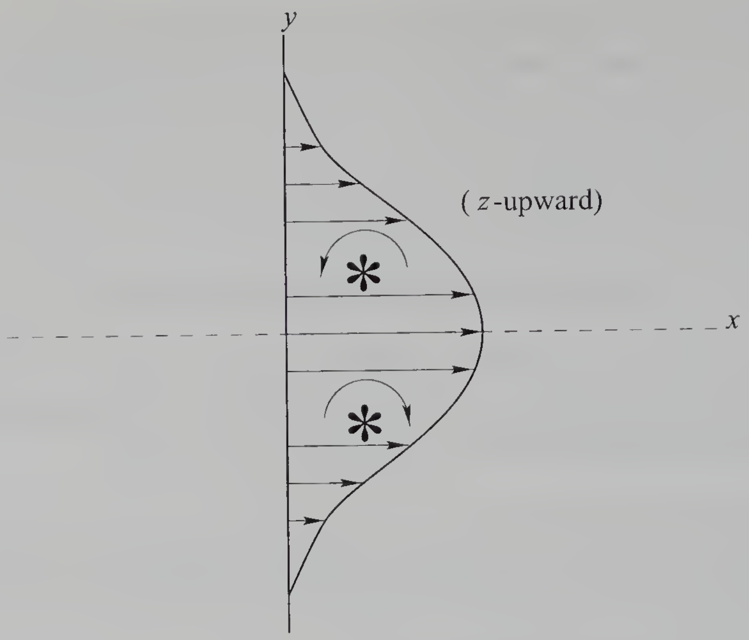

Fig. 7.4 Sketch of the flow field for a shear flow.

integral along the small contour. If the flow (locally) rotates and if we integrate along the fluid flow, the line integral $\oint \mathbf{v} \cdot d\mathbf{r}$ will be relatively large, so that this line integral indeed measures the local rotation.

Rotation and shear each contribute to the *curl* of a vector field. Let us consider once again a vector field of the form (7.6) which is axially symmetric around the z-axis. In the following we do not require the rotation around the z-axis to be rigid, so that $v(r)$ in (7.6) is arbitrary. We know that both the rotation around the z-axis and the shear are a source of vorticity.

Problem d Show that for the flow

$$v(r) = \frac{A}{r} \tag{7.11}$$

the vorticity vanishes, with A a constant that is not yet determined. Make a sketch of this flow field.

The vorticity of this flow vanishes despite the fact that the flow rotates around the z-axis (but not in rigid rotation) and that the flow has a nonzero shear. The reason that the vorticity vanishes is that the contribution of the rotation around the z-axis to the vorticity is equal but of opposite sign to the contribution of the shear, so that the total vorticity vanishes. Note that this implies that a paddle-wheel does not change its orientation as it moves with this flow!

7.5 Magnetic field induced by a straight current

At this point you may have the impression that the flow field (7.11) has been contrived in an artificial way. However, keep in mind that all the arguments of the previous section apply to any vector field and that fluid flow was used only as an example to fix our mind. As an example we consider the generation of the magnetic field **B** by an electrical current **J** that is independent of time. The Maxwell equation for the curl of the magnetic field in vacuum is for time-independent fields given by:

$$\nabla \times \mathbf{B} = \mu_0 \mathbf{J}, \tag{7.12}$$

see equation (5.22) in reference [53]. In this expression μ_0 is the magnetic permeability of vacuum. It plays the role of a coupling constant since it governs the strength of the magnetic field that is generated by a given current. It plays the same role as 1/permittivity in (6.23) or the gravitational constant G in (6.33). The vector **J** denotes the electric current per unit volume (properly called the electric current density).

For simplicity we consider an electric current running through an infinite straight wire along the z-axis. Because of rotational symmetry around the z-axis and because of translational invariance along the z-axis the magnetic field depends neither on φ nor on z and must be of the form (7.6). Away from the wire the electrical current **J** vanishes.

Problem a Use the results of problem b of the previous section to show that

$$\mathbf{B} = \frac{A}{r}\hat{\boldsymbol{\varphi}}. \tag{7.13}$$

A comparison with equation (7.11) shows that for this magnetic field the contribution of the "rotation" around the z-axis to $\nabla \times \mathbf{B}$ is exactly balanced by the contribution of the "magnetic shear" to $\nabla \times \mathbf{B}$. It should be noted that the magnetic field derived in this section is of great importance because this field has been used to define the unit of electrical current, the ampère. However, this can only be done when the constant A in expression (7.13) is known.

Problem b Why does the treatment of this section not tell us what the relation is between the constant A and the current **J** in the wire?

We return to this issue in Section 9.3.

7.6 Spherical coordinates and cylindrical coordinates

In Section 6.4, expressions for the divergence in spherical coordinates and cylindrical coordinates were derived. Here we do the same for the curl because these expressions are frequently very useful. It is possible to derive the curl in curvilinear coordinates by systematically transforming all the elements of the involved vectors and all the differentiations from Cartesian to curvilinear coordinates. As an alternative, we use the physical interpretation of the *curl* given by expression (7.3) to derive the *curl* in spherical coordinates. This expression states that a certain component of the *curl* of a vector field **v** is the line integral $\oint \mathbf{v} \cdot d\mathbf{r}$ along a contour perpendicular to the component of the *curl* that we are considering, normalized by the surface area bounded by that contour. As an example we derive for a system of spherical coordinates the φ-component of the *curl*; see Figure 7.5 for the definition of the geometric variables.

Consider the grey infinitesimal surface in Figure 7.5. When we carry out the line integration along the surface we integrate in the direction shown in the figure. The reason for this is that the azimuth φ increases when we move into the figure, hence $\hat{\boldsymbol{\varphi}}$ points into the figure. Following the rules of a right-handed screw, this

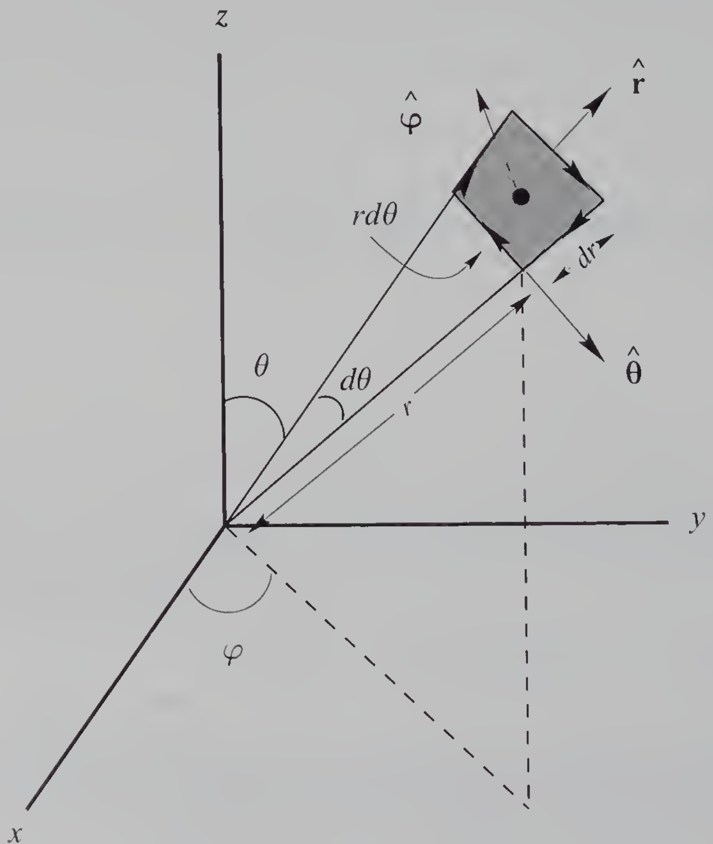

Fig. 7.5 Definition of the geometric variables for the computation of the *curl* in spherical coordinates.

corresponds to the indicated sense of integration. The area enclosed by the contour is given by $rd\theta dr$. By summing the contributions of the four sides of the contour we find, using expression (7.3), that the φ-component of $\nabla \times \mathbf{v}$ is given by:

$$(\nabla \times \mathbf{v})_\varphi = \frac{1}{rd\theta dr}[v_\theta(r+dr,\theta)(r+dr)\,d\theta - v_r(r,\theta+d\theta)\,dr \\ - v_\theta(r,\theta)r d\theta + v_r(r,\theta)\,dr]. \tag{7.14}$$

In this expression v_r and v_θ denote the components of $\mathbf{v}$ in the radial direction and in the direction of $\hat{\boldsymbol{\theta}}$ respectively.

Problem a Make sure that expression (7.14) is understood.

This result can be simplified by Taylor expanding the components of $\mathbf{v}$ in dr and $d\theta$ and linearizing the resulting expression in the infinitesimal increments dr and $d\theta$.

Problem b Do this and show that the final result does not depend on dr and $d\theta$ and is given by:

$$(\nabla \times \mathbf{v})_\varphi = \frac{1}{r}\frac{\partial}{\partial r}(r v_\theta) - \frac{1}{r}\frac{\partial v_r}{\partial \theta}. \tag{7.15}$$

The same treatment can be applied to the other components of the curl. This leads to the following expression for the curl in spherical coordinates:

$$\nabla \times \mathbf{v} = \hat{\mathbf{r}}\frac{1}{r\sin\theta}\left[\frac{\partial}{\partial\theta}\left(\sin\theta\; v_\varphi\right) - \frac{\partial v_\theta}{\partial\varphi}\right] + \hat{\boldsymbol{\theta}}\frac{1}{r}\left[\frac{1}{\sin\theta}\frac{\partial v_r}{\partial\varphi} - \frac{\partial}{\partial r}\left(r v_\varphi\right)\right] \\ + \hat{\boldsymbol{\varphi}}\frac{1}{r}\left[\frac{\partial}{\partial r}(r v_\theta) - \frac{\partial v_r}{\partial\theta}\right]. \tag{7.16}$$

Problem c Use a similar analysis to show that in cylindrical coordinates (r, φ, z) the curl is given by:

$$\nabla \times \mathbf{v} = \hat{\mathbf{r}}\left[\frac{1}{r}\frac{\partial v_z}{\partial\varphi} - \frac{\partial v_\varphi}{\partial z}\right] + \hat{\boldsymbol{\varphi}}\left[\frac{\partial v_r}{\partial z} - \frac{\partial v_z}{\partial r}\right] + \hat{\mathbf{z}}\frac{1}{r}\left[\frac{\partial}{\partial r}\left(r v_\varphi\right) - \frac{\partial v_r}{\partial\varphi}\right], \tag{7.17}$$

with $r = \sqrt{x^2 + y^2}$.

Problem d Use this result to re-derive (7.8) for vector fields of the form $\mathbf{v} = \mathrm{v}(r)\hat{\boldsymbol{\varphi}}$. Hint: use the same method as used in the derivation of (7.14) and treat the three components of the curl separately.

8

Theorem of Gauss

In Section 6.5 we determined the gravitational field in N dimensions using as the only ingredient that in free space, where the mass-density vanishes, the divergence of the gravitational field vanishes: $(\nabla \cdot \mathbf{g}) = 0$. This was sufficient to determine the gravitational field in expression (6.37). However, that expression is not quite satisfactory because it contains a constant A that is unknown. In fact, at this point we have no idea how this constant is related to the mass M that causes the gravitational field! The reason for this is simple: in order to derive the gravitational field in (6.37) we have only used the field equation (6.33) for free space (where $\rho = 0$). However, if we want to find the relation between the mass and the resulting gravitational field we must also use the field equation $(\nabla \cdot \mathbf{g}) = -4\pi G \rho$ at places where the mass is present. More specifically, we have to *integrate* the field equation in order to find the total effect of the mass. The theorem of Gauss gives us an expression for the volume integral of the divergence of a vector field.

8.1 Statement of Gauss's law

In Section 6.2 it was shown that the divergence is the flux per unit volume. In fact, expression (6.8) gives us the outward flux $d\Phi_v$ through an infinitesimal volume dV: $d\Phi_v = (\nabla \cdot \mathbf{v})dV$. We can integrate this expression to find the total flux through the surface S which encloses the total volume V:

$$\oint_S \mathbf{v} \cdot d\mathbf{S} = \int_V (\nabla \cdot \mathbf{v})\, dV. \tag{8.1}$$

In deriving this, equation (6.8) has been used to express the total flux on the left-hand side. This expression is called the theorem of Gauss.

Note that in the derivation of (8.1) we did not use the dimensionality of the space; this relation holds in any number of dimensions. You may recognize the one-dimensional version of (8.1). In one dimension the vector **v** has only one component v_x, hence $(\nabla \cdot \mathbf{v}) = \partial_x v_x$. A "volume" in one dimension is simply a line. Let this line run from $x = a$ to $x = b$. The "surface" of a one-dimensional volume consists of the endpoints of this line, so that the left-hand side of (8.1) is the difference of the function v_x at its endpoints. This implies that the theorem of Gauss in one-dimension is:

$$v_x(b) - v_x(a) = \int_a^b \frac{\partial v_x}{\partial x} dx. \tag{8.2}$$

This expression will be familiar to you. We will use the two-dimensional version of the theorem of Gauss in Section 9.2 to derive the theorem of Stokes.

Problem a Compute the flux of the vector field $\mathbf{v}(x, y, z) = (x + y + z)\hat{\mathbf{z}}$ through a sphere of radius R centered on the origin by explicitly computing the integral that defines the flux.

Problem b Solve problem a using Gauss's law (8.1).

8.2 Gravitational field of a spherically symmetric mass

In this section we use Gauss's law (8.1) to show that the gravitational field of a body with a spherically symmetric mass-density ρ depends only on the total mass and not on the distribution of the mass over that body. For a spherically symmetric body the mass-density depends only on radius: $\rho = \rho(r)$. Because of the spherical symmetry of the mass, the gravitational field is spherically symmetric and points in the radial direction:

$$\mathbf{g}(\mathbf{r}) = g(r)\hat{\mathbf{r}}. \tag{8.3}$$

Problem a Use the field equation (6.33) for the gravitational field and Gauss's law applied to a surface that completely encloses the mass to show that

$$\oint_S \mathbf{g} \cdot d\mathbf{S} = -4\pi G M, \tag{8.4}$$

where $M = \int_V \rho dV$ is the total mass of the body.

Problem b Use a sphere of radius r as the surface in (8.4) to show that the gravitational field is given in three dimensions by

$$\mathbf{g}(\mathbf{r}) = -\frac{GM}{r^2}\hat{\mathbf{r}}. \tag{8.5}$$

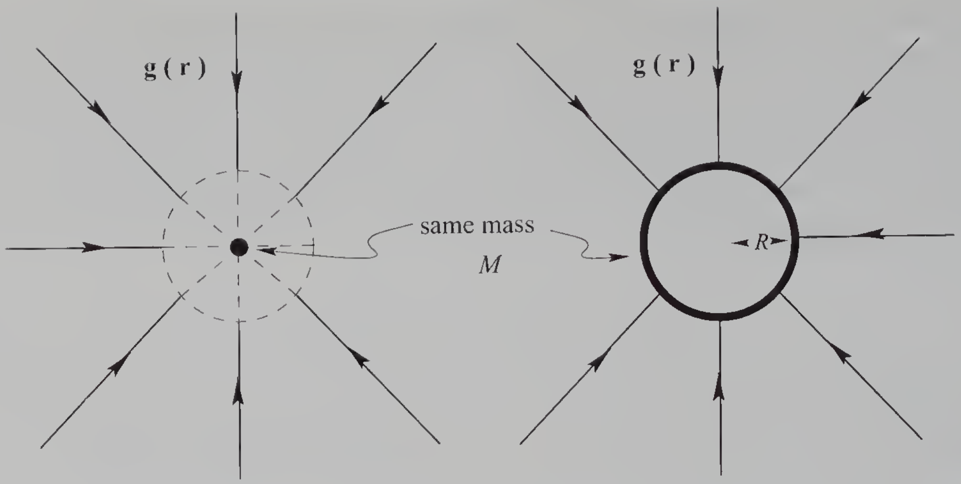

Fig. 8.1 Two different bodies with different mass distributions that generate the same gravitational field for distances larger than the radius of the body on the right.

This is an intriguing result. What we have shown here is that the gravitational field depends only on the total mass of the spherically symmetric body, but not on the distribution of the mass within that body. As an example consider two bodies with the same mass. One body has all its mass located in a small ball near the origin and the other body has all its mass distributed on a thin spherical shell of radius R, see Figure 8.1. According to (8.5) these bodies generate exactly the same gravitational field outside the body. This implies that gravitational measurements taken outside the two bodies cannot be used to distinguish between them. The nonunique relation between the gravity field and the underlying mass distribution is of importance for the interpretation of gravitational measurements taken in geophysical surveys.

Problem c Let us assume that the mass is located within a sphere of radius R, and that the mass-density within that sphere is constant. Integrate (6.33) over a sphere of radius $r < R$ to show that the gravitational field within the sphere is given by:

$$\mathbf{g}(\mathbf{r}) = -\frac{MGr}{R^3}\hat{\mathbf{r}}. \tag{8.6}$$

Plot the gravitational field as a function of r when the distance increases from zero to a distance larger than the radius R, where the gravitational field is given by equation (8.5). Verify that the gravitational field is continuous at the radius R of the sphere.

Note that all these conclusions hold identically for the electrical field when we replace the mass-density by the charge-density, because (6.23) for the divergence of the electric field has the same form as (6.33) for the gravitational field. As an example we consider a hollow spherical shell of radius R. On the spherical shell electrical charge is distributed with a constant charge density: $\rho = const.$

Problem d Use expression (6.23) for the electric field and Gauss's law to show that within the hollow sphere the electric field vanishes: $\mathbf{E}(\mathbf{r}) = 0$ for $r < R$.

This result implies that when a charge is placed within such a spherical shell the electrical field generated by the charge distribution on the shell exerts no net force on this charge; the charge will not move. Since the electrical potential satisfies $\mathbf{E} = -\nabla V$, the result derived in problem d implies that the potential is constant within the sphere. This property has actually been used to determine experimentally whether the electric field indeed satisfies (6.23) (which implies that the field of point charge decays as $1/r^2$). Measurements of the potential differences within a hollow spherical shell as described in problem d can be carried out with great sensitivity. Experiments based on this principle (usually in a more elaborate form) have been used to ascertain the decay of the electric field of a point charge with distance. Writing the field strength as $1/r^{2+\varepsilon}$ it has been shown that $\varepsilon = (2.7 \pm 3.1) \times 10^{-16}$, see Section I.2 of Jackson [53] for a discussion. The small value of ε is a remarkable experimental confirmation of (6.23) for the electric field.

8.3 Representation theorem for acoustic waves

Acoustic waves are waves that propagate through a gas or fluid. You can hear the voice of others because acoustic waves propagate from their vocal tract to your ear. Acoustic waves are frequently used to describe the propagation of waves through the Earth. Since the Earth is a solid body, this is strictly speaking not correct, but under certain conditions (small scattering angles) the errors can be acceptable. The pressure field $p(\mathbf{r})$ of acoustic waves satisfies the following partial differential equation in the frequency domain:

$$\nabla \cdot \left(\frac{1}{\rho} \nabla p \right) + \frac{\omega^2}{\kappa} p = f. \tag{8.7}$$

In this expression $\rho(\mathbf{r})$ is the mass-density of the medium, ω is the angular frequency and $\kappa(\mathbf{r})$ is the compressibility (a factor that describes how strongly the medium resists changes in its volume). The right-hand side $f(\mathbf{r})$ describes the source of the acoustic wave. This term accounts, for example, for the action of your voice.

We now consider two pressure fields $p_1(\mathbf{r})$ and $p_2(\mathbf{r})$ that both satisfy (8.7) with sources $f_1(\mathbf{r})$ and $f_2(\mathbf{r})$ on the right-hand side of the equation.

Problem a Multiply equation (8.7) for p_1 with p_2, multiply equation (8.7) for p_2 with p_1 and subtract the resulting expressions. Integrate the result over a volume V to show that:

$$\int_V \left[p_2 \nabla \cdot \left(\frac{1}{\rho} \nabla p_1 \right) - p_1 \nabla \cdot \left(\frac{1}{\rho} \nabla p_2 \right) \right] dV = \int_V (p_2 f_1 - p_1 f_2)\, dV. \quad (8.8)$$

Ultimately we want to relate the wave field at the surface S that encloses the volume V to the wave field within the volume. Obviously, Gauss's law is the tool for doing this. The problem we face is that Gauss's law holds for the volume integral of the divergence, whereas in (8.8) we have the *product* of a divergence (such as $\nabla \cdot \left(\rho^{-1} \nabla p_1 \right)$) with another function (such as p_2). This means we have to "make" a divergence.

Problem b First derive that

$$\nabla \cdot (f \mathbf{v}) = f (\nabla \cdot \mathbf{v}) + \mathbf{v} \cdot \nabla f. \quad (8.9)$$

Hint: write the left-hand side in its components and use the product rule for differentiation.

Problem c Use $f = p_2$ and $\mathbf{v} = \rho^{-1} \nabla p_1$ in (8.9) to show that

$$p_2 \nabla \cdot \left(\frac{1}{\rho} \nabla p_1 \right) = \nabla \cdot \left(\frac{1}{\rho} p_2 \nabla p_1 \right) - \frac{1}{\rho} (\nabla p_1 \cdot \nabla p_2). \quad (8.10)$$

What we are doing here is similar to the standard derivation of integration by parts. The easiest way to show that $\int_a^b f(\partial g/\partial x) dx = [f(x)g(x)]_a^b - \int_a^b (\partial f/\partial x) g dx$ is to integrate the identity $f(\partial g/\partial x) = \partial(fg)/dx - (\partial f/\partial x) g$ from $x = a$ to $x = b$. This last equation has exactly the same structure as (8.10).

Problem d Use (8.8), (8.10), and Gauss's law to derive that

$$\oint_S \frac{1}{\rho} (p_2 \nabla p_1 - p_1 \nabla p_2) \cdot d\mathbf{S} = \int_V (p_2 f_1 - p_1 f_2)\, dV. \quad (8.11)$$

To see the power of this expression, consider the special case that the source f_2 of p_2 is of unit strength and that this source is localized in a very small volume around a point $\mathbf{r}_0$ within the volume. This means that f_2 on the right-hand side of (8.11) is only nonzero at $\mathbf{r}_0$. The corresponding volume integral $\int_V p_1 f_2 dV$ is in that case given by $p_1(\mathbf{r}_0)$. The wavefield $p_2(\mathbf{r})$ generated by this point source is called the *Green's function*, and this special solution is denoted by $G(\mathbf{r}, \mathbf{r}_0)$. (The concept of the Green's function is introduced in great detail in Chapter 18.) The argument $\mathbf{r}_0$ is added to indicate that this is the wavefield at location $\mathbf{r}$ due to a unit source at location $\mathbf{r}_0$. We now consider a solution p_1 that has no sources within the volume V (i.e. $f_1 = 0$). Let us simplify the notation further by dropping the subscript "1" in p_1.

Problem e Show by making all these changes that (8.11) can be written as:

$$p(\mathbf{r}_0) = \oint_S \frac{1}{\rho} \left[p(\mathbf{r}) \nabla G(\mathbf{r}, \mathbf{r}_0) - G(\mathbf{r}, \mathbf{r}_0) \nabla p(\mathbf{r}) \right] \cdot d\mathbf{S}. \tag{8.12}$$

This result is called the "representation theorem" because it gives the wavefield inside the volume when the wavefield (and its gradient) are specified on the surface that bounds this volume. Expression (8.12) can be used to formally derive Huygens's principle which states that every point on a wavefront acts as a source for other waves and that interference between these waves determines the propagation of the wavefront. Equation (8.12) also forms the basis for imaging techniques for seismic data [93]. In seismic exploration one records the wavefield at the Earth's surface. This can be used by taking the Earth's surface as the surface S over which the integration is carried out. If the Green's function $G(\mathbf{r}, \mathbf{r}_0)$ is known, one can use expression (8.12) to compute the wavefield in the interior of the Earth. Once the wavefield in the interior of the Earth is known, one can deduce some of the properties of the material in the Earth. In this way, equation (8.12) (or its generalization to elastic waves [3]) forms the basis of seismic imaging techniques.

This treatment suggests that if one knows the wavefield at a surface, one can determine the wavefield anywhere within the surface, and hence one can determine the properties of the medium within the surface. This in turn suggests that the problem of seismic imaging has been solved by the treatment of this section. There is, however, a catch. In order to apply (8.12) one must know the Green's function G. To know this function one must know the medium. We thus have the interesting situation that once we know the Green's function, we can deduce the properties of the medium from (8.12), but that in order to use the Green's function we must know the medium. This suggests that seismic imaging cannot be carried out. Fortunately, it turns out that in practice it suffices to use a reasonable *estimate* of the Green's function in (8.12). Such an estimated Green's function can be computed once the velocity in the medium is reasonably well known. It is for this reason that *velocity estimation* is such a crucial step in seismic data processing [28, 123].

8.4 Flowing probability

In classical mechanics, the motion of a particle with mass m is governed by Newton's law: $m\ddot{\mathbf{r}} = \mathbf{F}$. When the force $\mathbf{F}$ is associated with a potential $V(\mathbf{r})$ the motion of the particle satisfies:

$$m\frac{d^2\mathbf{r}}{dt^2} = -\nabla V(\mathbf{r}). \tag{8.13}$$

Newton's law, however, does not hold for particles that are very small. Atomic particles such as electrons are not described accurately by (8.13). One of the outstanding features of quantum mechanics is that an atomic particle is treated as a wave that describes the properties of that particle. This rather vague statement reflects the wave-particle duality that forms the basis of quantum mechanics. The wavefunction $\psi(\mathbf{r},t)$ that describes a particle that moves under the influence of a potential $V(\mathbf{r})$ satisfies the Schrödinger equation:[‡]

$$i\hbar\frac{\partial\psi(\mathbf{r},t)}{\partial t} = -\frac{\hbar^2}{2m}\nabla^2\psi(\mathbf{r},t) + V(\mathbf{r})\psi(\mathbf{r},t). \tag{8.14}$$

In this expression, $\hbar$ is Planck's constant h divided by 2π; where Planck's constant has the numerical value $h = 6.626 \times 10^{-34}$ kg m^2/s.

Suppose we are willing to accept that the motion of an electron is described by the Schrödinger equation, then the following question arises: What is the position of the electron as a function of time? According to the Copenhagen interpretation of quantum mechanics, this is a meaningless question because the electron behaves like a wave and does not have a definite location. Instead, the wavefunction $\psi(\mathbf{r},t)$ dictates how likely it is that the particle is at location $\mathbf{r}$ at time t. Specifically, the quantity $|\psi(\mathbf{r},t)|^2$ is the probability-density of finding the particle at location $\mathbf{r}$ at time t. This implies that the probability P_V that the particle is located within the volume V is given by $P_V = \int_V |\psi|^2 dV$. (Take care not to confuse the volume with the potential, because they are both indicated with the same symbol V.) This implies that the wavefunction is related to a probability. Instead of the motion of the electron, Schrödinger's equation dictates how the probability-density of the particle moves through space as time progresses. One expects that a "probability current" is associated with this movement. In this section we determine this current using the theorem of Gauss.

Problem a In the following we need the time derivative of $\psi^*(\mathbf{r},t)$, where the asterisk denotes the complex conjugate. Derive the differential equation that $\psi^*(\mathbf{r},t)$ obeys by taking the complex conjugate of Schrödinger's equation (8.14).

[‡] In this expression ∇^2 stands for the Laplacian which is treated in Chapter 10. At this point you only need to know that the Laplacian of a function is the divergence of the gradient of that function: $\nabla^2\psi = \text{div grad }\psi = \nabla\cdot\nabla\psi$.

Problem b Use this result to derive that for a volume V that is fixed in time:

$$\frac{\partial}{\partial t}\int_V |\psi|^2 dV = \frac{i\hbar}{2m}\int_V (\psi^*\nabla^2\psi - \psi\nabla^2\psi^*)dV. \tag{8.15}$$

Hint: use that $\frac{\partial}{\partial t}|\psi|^2 = \frac{\partial}{\partial t}(\psi\psi^*) = \psi\frac{\partial\psi^*}{\partial t} + \psi^*\frac{\partial\psi}{\partial t}$, and use (8.14) and the result of problem a.

Problem c Use Gauss's law to rewrite this expression as:

$$\frac{\partial}{\partial t}\int_V |\psi|^2 dV = \frac{i\hbar}{2m}\oint(\psi^*\nabla\psi - \psi\nabla\psi^*)\cdot d\mathbf{S}. \tag{8.16}$$

Hint: use a treatment similar to expression (8.10).

The left-hand side of this expression gives the time derivative of the probability that the particle is within the volume V. The only way the particle can enter or leave the volume is through the enclosing surface S. The right-hand side therefore describes the "flow" of probability through the surface S. More accurately, one can formulate this as the flux of the probability-density current. This means that expression (8.16) can be written as

$$\frac{\partial P_V}{\partial t} = -\int \mathbf{J}\cdot d\mathbf{S}, \tag{8.17}$$

where the probability-density current $\mathbf{J}$ is given by:

$$\mathbf{J} = \frac{i\hbar}{2m}(\psi\nabla\psi^* - \psi^*\nabla\psi). \tag{8.18}$$

As an example let us consider a plane wave:

$$\psi(\mathbf{r}, t) = A\, e^{i(\mathbf{k}\cdot\mathbf{r} - \omega t)}, \tag{8.19}$$

where $\mathbf{k}$ is the wavevector and A an unspecified constant.

Problem d Show that the wavelength λ is related to the wavevector by the relation $\lambda = 2\pi/|\mathbf{k}|$. In which direction does the wave propagate?

Problem e Show that the probability-density current $\mathbf{J}$ for this wavefunction satisfies:

$$\mathbf{J} = \frac{\hbar\mathbf{k}}{m}|\psi|^2. \tag{8.20}$$

This is a very interesting expression. The term $|\psi|^2$ gives the probability-density of the particle, while the probability-density current $\mathbf{J}$ physically describes the current of this probability-density. Since the probability-density current moves with the velocity of the particle, the remaining terms on the right-hand side of (8.20) must denote the velocity of the particle:

$$\mathbf{v} = \frac{\hbar \mathbf{k}}{m}. \tag{8.21}$$

Since the momentum $\mathbf{p}$ is the mass times the velocity, (8.21) can also be written as $\mathbf{p} = \hbar\mathbf{k}$. This relation was proposed by de Broglie in 1924 using completely different arguments than those we have used here [22]. Its discovery was a major step in the development of quantum mechanics.

Problem f Use this expression and the result of problem e to compute your own wavelength while you are riding your bicycle. Are quantum-mechanical phenomena important when you ride your bicycle? Use your wavelength as an argument. Did you know you possessed a wavelength?

9

Theorem of Stokes

In Chapter 8 we noted that in order to find the gravitational field of a mass we have to integrate the field equation (6.33) over the mass. Gauss's theorem can then be used to compute the integral of the divergence of the gravitational field. For the curl the situation is similar. In Section 7.5 we computed the magnetic field generated by a current in a straight infinite wire. The field equation

$$\nabla \times \mathbf{B} = \mu_0 \mathbf{J} \tag{7.12}$$

was used to compute the field away from the wire. However, the solution (7.13) contained an unknown constant A. The reason for this is that the field equation (7.12) was only used outside the wire, where $\mathbf{J} = 0$. The treatment of Section 7.5 therefore did not provide us with the relation between the field $\mathbf{B}$ and its source $\mathbf{J}$. The only way to obtain this relation is to integrate the field equation. This implies we have to compute the integral of the curl of a vector field. The theorem of Stokes tells us how to do this.

9.1 Statement of Stokes's law

The theorem of Stokes is based on the principle that the curl of a vector field is the closed line integral of the vector field per unit surface area, see Section 7.1. Mathematically this statement is expressed by (7.2) which we write in a slightly different form as:

$$\oint_{d\mathbf{S}} \mathbf{v} \cdot d\mathbf{r} = (\boldsymbol{\nabla} \times \mathbf{v}) \cdot \hat{\mathbf{n}}\, dS = (\boldsymbol{\nabla} \times \mathbf{v}) \cdot d\mathbf{S}. \tag{9.1}$$

The only difference with (7.2) is that in the above expression we have not aligned the z-axis with the vector $\boldsymbol{\nabla} \times \mathbf{v}$. The infinitesimal surface is therefore not necessarily confined to the (x, y)-plane and the z-component of the curl is replaced by the component of the *curl* normal to the surface, hence the occurrence of the terms

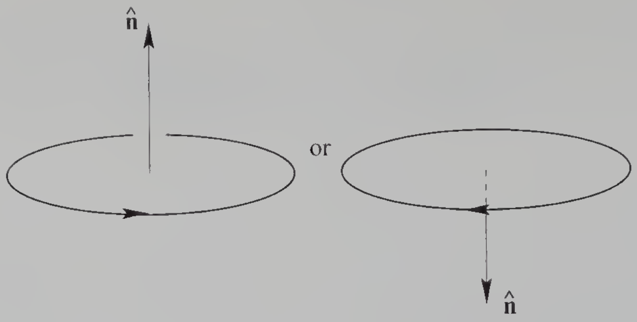

Fig. 9.1 The relation between the sense of integration and the orientation of the surface.

$\hat{\mathbf{n}}dS$ in (9.1). Expression (9.1) holds for an infinitesimal surface area. However, this expression can immediately be integrated to give the surface integral of the curl over a finite surface S that is bounded by the curve C:

$$\oint_C \mathbf{v} \cdot d\mathbf{r} = \int_S (\nabla \times \mathbf{v}) \cdot d\mathbf{S}. \tag{9.2}$$

This result is known as the theorem of Stokes (or Stokes's law). The line integral on the left-hand side is over the curve that bounds the surface S. A proper derivation of Stokes's law can be found in reference [68].

Note that a line integration along a closed surface can be carried out in two directions. What is the direction of the line integral on the left-hand side of Stokes's law (9.2)? To see this, we have to realize that Stokes's law was ultimately based on (7.2). The orientation of the line integration used in that expression is defined in Figure 7.1 , where it can be seen that the line integration is in the counterclockwise direction. In Figure 7.1 the z-axis points out of the paper, which implies that the vector $d\mathbf{S}$ also points out of the paper. *This means that in Stokes's law the sense of the line integration and the direction of the surface vector* $d\mathbf{S}$ *are related through the rule for a right-handed screw. This orientation is indicated in Figure* 9.1.

There is something strange about Stokes's law. If we define a curve C over which we carry out the line integration, we can define many different surfaces S that are bounded by the same curve C. Apparently, the surface integral on the right-hand side of Stokes's law does not depend on the specific choice of the surface S as long as it is bounded by the curve C.

Problem a Let us verify this property for an example. Consider the vector field $\mathbf{v} = r\hat{\boldsymbol{\varphi}}$. Let the curve C used for the line integral be a circle in the (x, y)-plane of radius R, see Figure 9.2 for the geometry of the problem. (i) Compute the line integral on the left-hand side of (9.2) by direct integration. Compute the surface integral on the right-hand side of (9.2) by (ii) integrating over

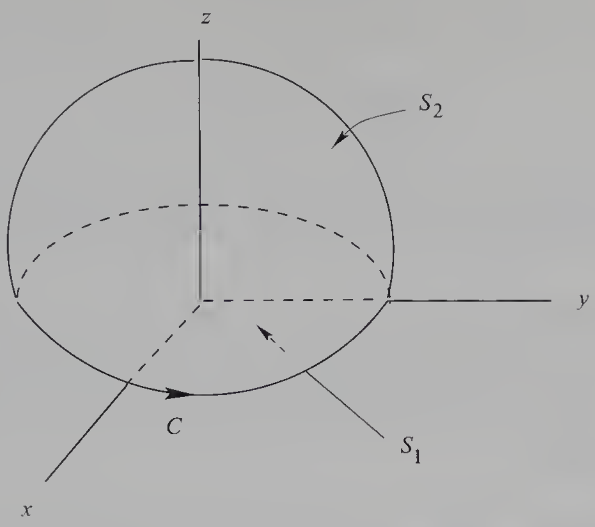

Fig. 9.2 Definition of the geometric variables for problem a.

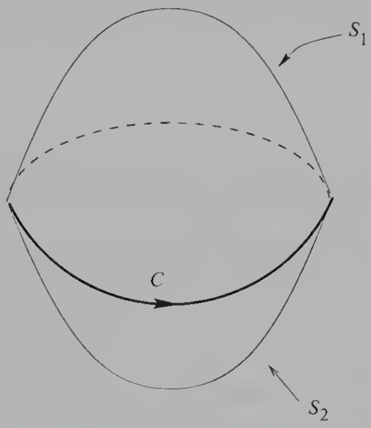

Fig. 9.3 Two surfaces that are bounded by the same contour C.

a disc of radius R in the (x, y)-plane (the surface S_1 in Figure 9.2) and by (iii) integrating over the upper half of a sphere with radius R (the surface S_2 in Figure 9.2). Verify that the three integrals are identical.

It is actually not difficult to prove that the surface integral in Stokes's law is independent of the specific choice of the surface S as long as it is bounded by the same contour C. Consider Figure 9.3 where the two surfaces S_1 and S_2 are bounded by the same contour C. We want to show that the surface integral of $\nabla \times \mathbf{v}$ is the same for the two surfaces, that is that:

$$\int_{S_1} (\nabla \times \mathbf{v}) \cdot d\mathbf{S} = \int_{S_2} (\nabla \times \mathbf{v}) \cdot d\mathbf{S}. \tag{9.3}$$

We can form a closed surface S by combining the surfaces S_1 and S_2.

Problem b Show that (9.3) is equivalent to the condition

$$\oint_S (\nabla \times \mathbf{v}) \cdot d\mathbf{S} = 0, \tag{9.4}$$

where the integration is over the closed surfaces defined by the combination of S_1 and S_2. Pay particular attention to the signs of the different terms.

Problem c Use Gauss's law to convert (9.4) to a volume integral and show that the integral is indeed identical to zero. In doing so you need to use the identity $\nabla \cdot (\nabla \times \mathbf{v}) = 0$ (or in another notation div curl $\mathbf{v} = 0$). Make sure you can derive this identity.

The result you obtained in problem c implies that condition (9.3) is indeed satisfied and that in the application of Stokes's law you can choose *any* surface as long as it is bounded by the contour over which the line integration is carried out. This is a very useful result because often the surface integration can be simplified by choosing the surface carefully.

9.2 Stokes's theorem from the theorem of Gauss

Stokes's law is concerned with surface integrations. Since the curl is intrinsically a three-dimensional vector, Stokes's law is inherently related to three space dimensions. However, if we consider a vector field that depends only on the coordinates x and y ($\mathbf{v} = \mathbf{v}(x, y)$) and that has a vanishing component in the z-direction ($v_z = 0$), then $\nabla \times \mathbf{v}$ points along the z-axis. If we consider a contour C that is confined to the (x, y)-plane, for such a vector field Stokes's law (9.2) takes the form

$$\oint_C (v_x dx + v_y dy) = \int_S \left(\partial_x v_y - \partial_y v_x\right) dxdy. \tag{9.5}$$

Problem a Verify this.

This result can be derived from the theorem of Gauss in two dimensions.

Problem b Show that Gauss's law (8.1) for a vector field $\mathbf{u}$ in two dimensions can be written as

$$\oint_C (\mathbf{u} \cdot \hat{\mathbf{n}}) ds = \int_S (\partial_x u_x + \partial_y u_y) dxdy, \tag{9.6}$$

where the unit vector $\hat{\mathbf{n}}$ is perpendicular to the curve C (see Figure 9.4) and where ds denotes the integration over the arclength of the curve C.

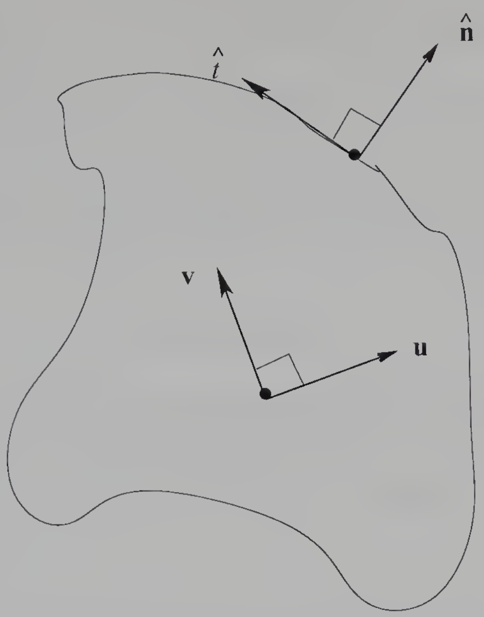

Fig. 9.4 Definition of the geometric variables for the derivation of Stokes's law from the theorem of Gauss.

In order to derive the special form of Stokes's law (9.5) from Gauss's law (9.6) we have to define the relation between the vectors **u** and **v**. Let the vector **u** follow from **v** by a clockwise rotation over 90 degrees, see Figure 9.4.

Problem c Show that:

$$v_x = -u_y \qquad \text{and} \qquad v_y = u_x . \tag{9.7}$$

We now define the unit vector $\hat{\mathbf{t}}$ to be directed along the curve C, see Figure 9.4. Since a rotation is an orthonormal transformation, the inner product of two vectors is invariant for a rotation over 90 degrees so that $(\mathbf{u} \cdot \hat{\mathbf{n}}) = (\mathbf{v} \cdot \hat{\mathbf{t}})$.

Problem d Verify this by expressing the components of $\hat{\mathbf{t}}$ in the components of $\hat{\mathbf{n}}$ and by using (9.7). The change in the position vector along the curve C is given by

$$d\mathbf{r} = \hat{\mathbf{t}}\, ds = \begin{pmatrix} dx \\ dy \end{pmatrix} . \tag{9.8}$$

Problem e Use these results to show that (9.5) follows from (9.6).

What you have shown here is that Stokes's law for the special case considered in this section is identical to the theorem of Gauss for two spatial dimensions.

9.3 Magnetic field of a current in a straight wire

We now return to the problem of the generation of the magnetic field induced by a current in an infinite straight wire that was discussed in Section 7.5. Because of the cylindrical symmetry of the problem, we know that the magnetic field is oriented in the direction of the unit vector $\hat{\varphi}$ and that the field only depends on the distance $r = \sqrt{x^2 + y^2}$ to the wire:

$$\mathbf{B} = B(r)\hat{\varphi}. \tag{9.9}$$

The field can be found by integrating the field equation $\nabla \times \mathbf{B} = \mu_0 \mathbf{J}$ over a disc of radius r perpendicular to the wire, see Figure 9.5. When the disc is larger than the thickness of the wire the surface integral of $\mathbf{J}$ gives the electric current I through the wire: $I = \int \mathbf{J} \cdot d\mathbf{S}$.

Problem a Use these results and Stokes's law to show that:

$$\mathbf{B} = \frac{\mu_0 I}{2\pi r}\hat{\varphi}. \tag{9.10}$$

We now have a relation between the magnetic field and the current that generates the field, hence the constant A in expression (7.13) is now determined. Note that the magnetic field depends only on the total current through the wire: it does not depend on the distribution of the electric current density $\mathbf{J}$ within the wire as long as the electric current-density exhibits cylindrical symmetry. Compare this with the result you obtained in problem b of Section 8.2 for the gravitational field!

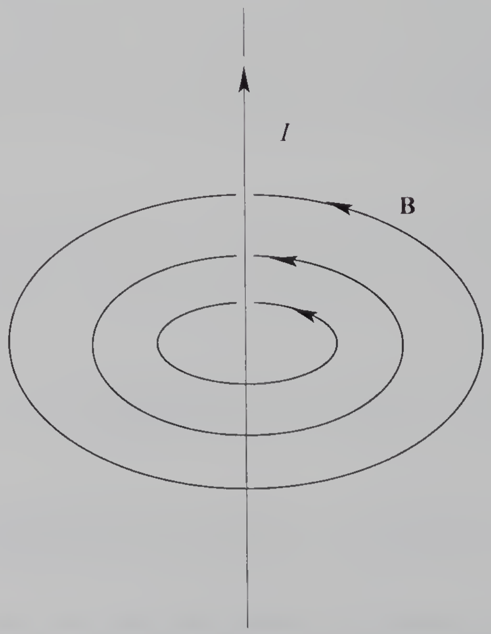

Fig. 9.5 Geometry of the magnetic field induced by a current in a straight infinite wire.

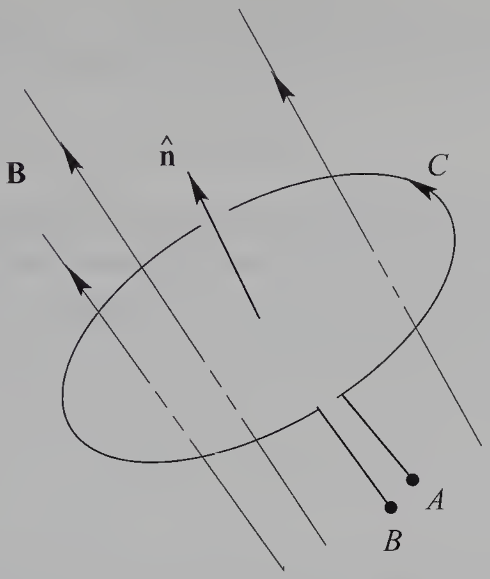

Fig. 9.6 A wire-loop in a time-dependent magnetic field.

9.4 Magnetic induction and Lenz's law

The theory in the previous section deals with the generation of a magnetic field by a current. A magnet placed in this field will experience a force exerted by the magnetic field. This force is essentially the driving force in electric motors; using an electrical current that changes with time a time-dependent magnetic field is generated that exerts a force on magnets attached to a rotation axis.

In this section we study the reverse effect: what is the electrical field generated by a magnetic field that changes with time? In a dynamo, a moving part (e.g. a bicycle wheel) moves a magnet. This creates a time-dependent electric field. This process is called magnetic induction and is described by the following Maxwell equation (see reference [53]):

$$\nabla \times \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t}. \tag{9.11}$$

To fix our mind, let us consider a wire with endpoints A and B, see Figure 9.6. The direction of the magnetic field is indicated in this figure. In order to find the electric field induced in the wire, integrate (9.11) over the surface enclosed by the wire[‡]

$$\int_S (\nabla \times \mathbf{E}) \cdot d\mathbf{S} = -\int_S \frac{\partial \mathbf{B}}{\partial t} \cdot d\mathbf{S}. \tag{9.12}$$

[‡] You may feel uncomfortable applying Stokes's law to the open curve AB. However, remember that the electric field associated with the varying magnetic field is a continuous function of the space variables. The electric field is therefore the same at the points A and B where the wire is open. The line integral over the open curve C is therefore identical to the line integral along the closed contour.

Problem a Show that the right-hand side of (9.12) is given by $-\partial\Phi/\partial t$, where Φ is the magnetic flux through the wire. (See Section 6.1 for the definition of the flux.)

We have discovered that a change in the magnetic flux is the source of an electric field. The resulting field can be characterized by the *electromotive force* F_{AB}, which is a measure of the work done by the electric field on a unit charge when it moves from point A to point B, see Figure 9.6:

$$F_{AB} \equiv \int_A^B \mathbf{E} \cdot d\mathbf{r}. \tag{9.13}$$

Problem b Apply Stokes's law to the left-hand side of expression (9.12) to show that the electromotive force satisfies

$$F_{AB} = -\frac{\partial \Phi_{\mathbf{B}}}{\partial t}. \tag{9.14}$$

Problem c Because of the electromotive force an electric current will flow through the wire. Determine the direction of the electric current in the wire. Show that this current generates a magnetic field that opposes the change in the magnetic field that generates the current. You learned in Section 9.3 the direction of the magnetic field that is generated by an electric current in a wire.

What we have discovered in problem c is Lenz's law, which states that induction currents lead to a secondary magnetic field which opposes the change in the primary magnetic field that generates the electric current. This implies that coils in electrical systems exhibit a certain inertia in the sense that they resist changes in the magnetic field that passes through the coil. The amount of inertia is described by a quantity called the inductance L. This quantity plays a role similar to mass in classical mechanics because the mass of a body also describes how strongly a body resists changing its velocity when an external force is applied.

9.5 Aharonov–Bohm effect

It was shown in Section 6.3 that because of the absence of magnetic monopoles the magnetic field is source-free: $(\nabla \cdot \mathbf{B}) = 0$. In electromagnetism one often expresses the magnetic field as the curl of a vector field $\mathbf{A}$:

$$\mathbf{B} = \nabla \times \mathbf{A}. \tag{9.15}$$

The advantage of writing the magnetic field in this way is that for *any* field $\mathbf{A}$ the magnetic field satisfies $(\nabla \cdot \mathbf{B}) = 0$ because $\nabla \cdot (\nabla \times \mathbf{A}) = 0$.

Problem a Give a proof of this last identity.

The vector field $\mathbf{A}$ is called the *vector potential*. The reason for this name is that it plays a role similar to the electric potential V. The electric and the magnetic fields follow from V and $\mathbf{A}$ respectively by differentiation: $\mathbf{E} = -\nabla V$ and $\mathbf{B} = \nabla \times \mathbf{A}$. The electric field vanishes when the potential V is constant. Since this constant may be nonzero, this implies that the electric field may vanish but that the potential is nonzero. The vector potential has a similar property, that is it can be nonzero (and variable) in parts of space where the magnetic field vanishes. As an example, consider a magnetic field with cylindrical symmetry along the z-axis which is constant for $r < R$ and which vanishes for $r > R$:

$$\mathbf{B} = \begin{cases} B_0\hat{\mathbf{z}} & \text{for} \quad r < R \\ 0 & \text{for} \quad r > R \end{cases}, \tag{9.16}$$

see Figure 9.7 for a sketch of the magnetic field. Because of the cylindrical symmetry the vector potential is a function of the distance r to the z-axis only and does not depend on z or φ.

Problem b Use expression (7.17) to show that a vector potential of the form

$$\mathbf{A} = f(r)\hat{\boldsymbol{\varphi}} \tag{9.17}$$

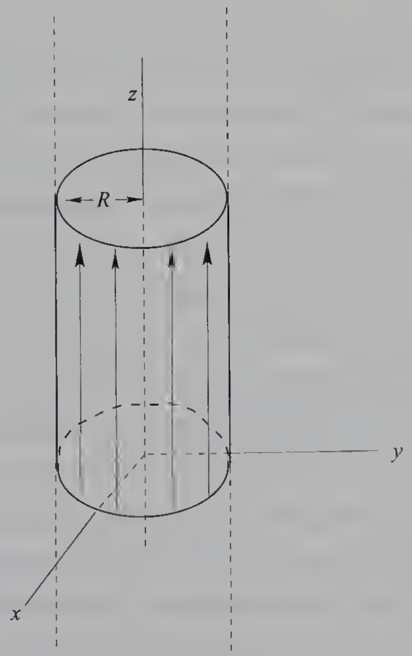

Fig. 9.7 Geometry of the magnetic field.

gives a magnetic field in the required direction. Give a derivation showing that $f(r)$ satisfies the following differential equation:

$$\frac{1}{r}\frac{\partial}{\partial r}(rf(r)) = \begin{cases} B_0 & \text{for} \quad r < R \\ 0 & \text{for} \quad r > R \end{cases} . \tag{9.18}$$

This differential equation for $f(r)$ can be immediately integrated. After integration, two integration constants are present. These constants follow from the requirements that the vector potential is continuous at $r = R$ and that $f(r = 0) = 0$. (This latter requirement is needed because the direction of the unit vector $\hat{\varphi}$ is undefined on the z-axis, where $r = 0$. The vector potential therefore only has a unique value at the z-axis when $f(r = 0) = 0$.)

Problem c Integrate the differential equation (9.18) subject to the boundary condition for $f(r = 0)$ to derive that the vector potential is given by

$$\mathbf{A} = \begin{cases} \frac{1}{2}B_0 r\,\hat{\varphi} & \text{for} \quad r < R \\ \frac{1}{2}B_0 \frac{R^2}{r}\hat{\varphi} & \text{for} \quad r > R \end{cases} . \tag{9.19}$$

The importance of this expression is that although the magnetic field is only nonzero for $r < R$, the vector potential (and its gradient) is nonzero everywhere in space! The vector potential is thus much more nonlocal than the magnetic field. This leads to a very interesting effect in quantum mechanics, called the Aharonov–Bohm effect.

Before introducing this effect we need to know more about quantum mechanics. As you have seen in Section 8.4, the behavior of atomic "particles" such as electrons is paradoxically described by a wave. The properties of this wave are described by Schrödinger's equation (8.14). When different waves propagate in the same region of space, interference can occur. In some parts of space the waves may enhance each other (constructive interference) while in other parts the waves cancel each other (destructive interference). This is observed for "particle waves" when electrons are sent through two slits and then detected on a screen behind these slits, see the left-hand panel of Figure 9.8. You might expect the electrons to propagate like bullets along straight lines. In that case, electrons would be detected only at the intersection of the screen with these straight lines. This is, however, not the case: in experiments one observes a pattern of fringes on the screen that is caused by the constructive and destructive interference of the electron waves. This interference pattern is sketched in Figure 9.8 on the right-hand side of the screens. This remarkable confirmation of the wave property of particles is described clearly in reference [39]. (The situation is

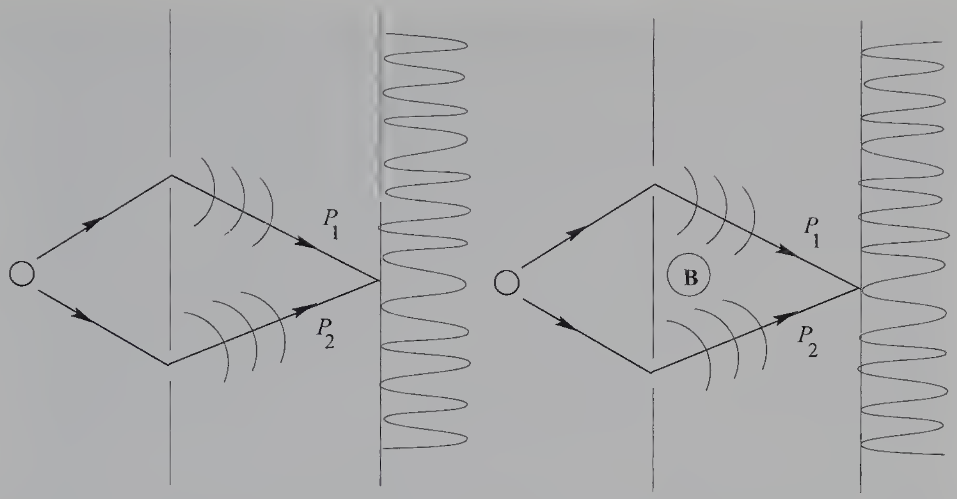

Fig. 9.8 Experiment in which electrons travel through two slits and are detected on a screen behind the slits. The resulting interference pattern is sketched. The experiment without a magnetic field is shown on the left, the experiment with a magnetic field is shown on the right. Note the shift in the maxima and minima of the interference pattern between the two experiments.

even more remarkable when one sends the electrons through the slits "one-by-one" so that only one electron passes through the slits at a time. In that case one sees a dot at the detector for each electron. However, after many particles have arrived at the detector this pattern of dots forms the interference pattern of the waves, see reference [94].)

Let us consider the same experiment, but with a magnetic field given by (9.16) placed between the two slits. When the electrons propagate along the paths P_1 or P_2 they do not pass through this field, hence one would expect that the electrons would not be influenced by this field and that the magnetic field would not change the observed interference pattern at the detector. However, it is an observational fact that the magnetic field *does* change the interference pattern at the detector, see reference [94] for examples. This surprising effect is called the Aharonov–Bohm effect [2].

In order to understand this effect, we should note that a magnetic field in quantum mechanics leads to a phase shift of the wavefunction. If the wavefunction in the absence of a magnetic field is given by $\psi(\mathbf{r})$, the wavefunction in the presence of the magnetic field is given by $\psi(\mathbf{r})\times \exp[(ie/\hbar c)\int_P \mathbf{A}\cdot d\mathbf{r}]$, see reference [91]. In this expression $\hbar$ is Planck's constant (divided by 2π), c is the speed of light, and $\mathbf{A}$ is the vector potential associated with the magnetic field. The integration is over the path P from the source of the particles to the detector. Consider now the waves that interfere in the two-slit experiment in the right-hand panel of Figure 9.8. The wave that travels through the upper slit experiences a phase shift $\exp[(ie/\hbar c)\int_{P_1} \mathbf{A}\cdot d\mathbf{r}]$, where the integration is over the path P_1 through the upper slit. The wave that travels through the lower slit obtains a phase shift $\exp[(ie/\hbar c)\int_{P_2}\mathbf{A}\cdot d\mathbf{r}]$ where the path P_2 runs through the lower slit.

Problem d Show that the phase difference $\delta\varphi$ between the two waves due to the presence of the magnetic field is given by

$$\delta\varphi = \frac{e}{\hbar c}\oint_P \mathbf{A}\cdot d\mathbf{r}, \tag{9.20}$$

where the path P is the closed path from the source through the upper slit to the detector and back through the lower slit to the source.

This phase difference affects the interference pattern because it is the *relative* phase between interfering waves that determines whether the interference is constructive or destructive.

Problem e Use Stokes's law and expression (9.15) to show that the phase difference can be written as

$$\delta\varphi = \frac{e\Phi_{\mathbf{B}}}{\hbar c}, \tag{9.21}$$

where $\Phi_{\mathbf{B}}$ is the magnetic flux through the area enclosed by the path P.

This expression shows that the phase shift between the interfering waves is proportional to the magnetic field enclosed by the paths of the interfering waves, *despite the fact that the electrons never move through the magnetic field* **B**. Mathematically the reason for this surprising effect is that the vector potential is nonzero throughout space even when the magnetic field is confined to a small region of space, see (9.19) as an example. However, this explanation is purely mathematical and does not seem to agree with common sense. This has led to speculations that the vector potential is actually a more "fundamental" quantity than the magnetic field [94].

9.6 Wingtips vortices

If you have watched aircraft closely, you may have noticed that sometimes a stream of condensation is left behind by the wingtips, see Figure 9.9. This is a different condensation trail than the thick contrails created by the engines. The condensation trails that start at the wingtips are due to a vortex (a spinning motion of the air) that is generated at the wingtips. This vortex is called the wingtip vortex. In this section we use Stokes's law to see that this wingtip vortex is closely related to the lift that is generated by the airflow along a wing.

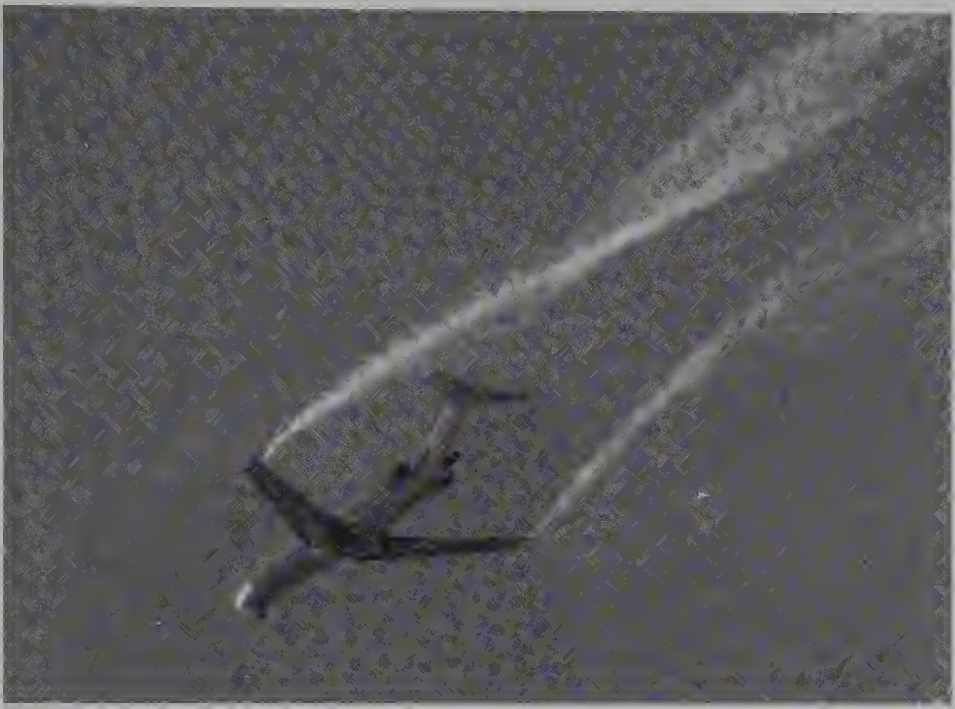

Fig. 9.9 Vortices trailing from the wingtips of a Boeing 727. Figure courtesy of NASA.

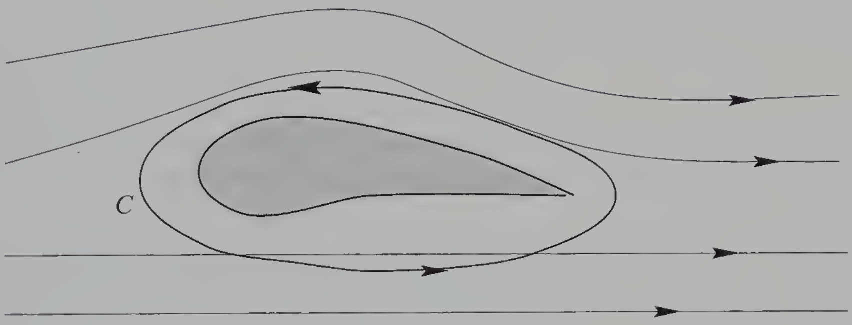

Fig. 9.10 Sketch of the flow along an airfoil. The wing is shown in grey, the contour C is shown by the thick solid line.

Let us first consider the flow along a wing, see Figure 9.10. In the figure the air traverses a longer path along the upper part of the wing than along the lower part. Because of the curved upper side of the wing, the velocity of the airstream along the upper part of the wing is larger than the velocity along the lower part.

Problem a The *circulation* is defined as the line integral $\oint_C \mathbf{v} \cdot d\mathbf{r}$ of the velocity along a curve. Is the circulation positive or negative for the curve C in Figure 9.10 for the indicated sense of integration? Use that the air moves faster over the upper side of the wing than under the lower side.

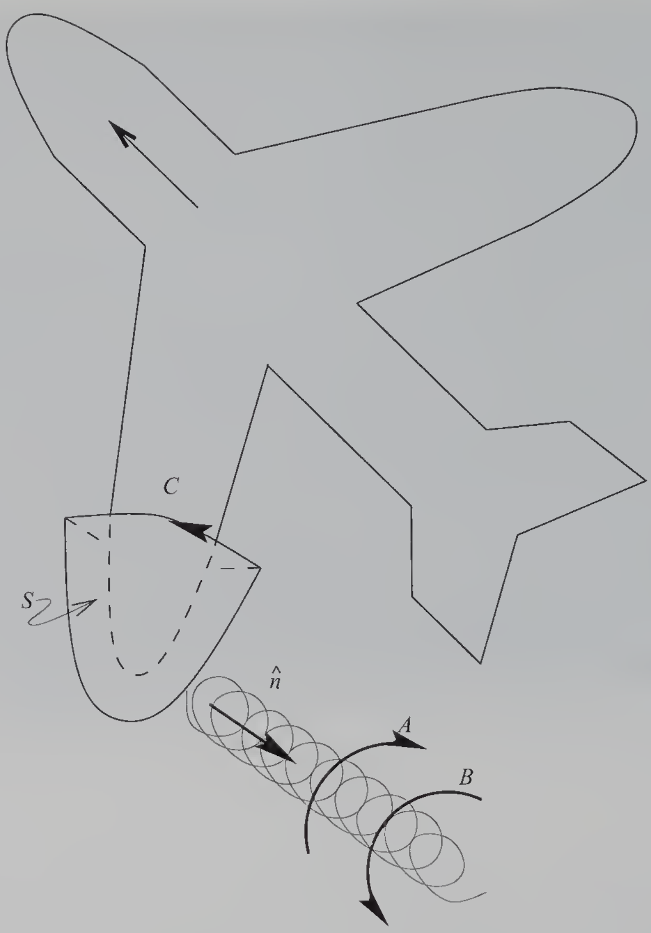

Fig. 9.11 Geometry of the surface S and the wingtip vortex for an aircraft seen from above. The surface S encloses the wingtip of the aircraft. The edge of this surface is the same contour C as drawn in the previous figure.

Problem b Consider now the surface S shown in Figure 9.11. Show that the circulation satisfies

$$\oint_C \mathbf{v} \cdot d\mathbf{r} = \int_S \omega \cdot d\mathbf{S}, \tag{9.22}$$

where ω is the vorticity. (See Sections 7.2–7.4.)

This expression implies that whenever lift is generated by the circulation along the contour C around the wing, the integral of the vorticity over a surface that envelopes the wingtip is nonzero. The vorticity depends on the derivative of the velocity. Since the flow is relatively smooth along the wing, the derivative of the velocity field is largest near the wingtips. Therefore, expression (9.22) implies that vorticity is generated at the wingtips. As shown in Section 7.3 the vorticity is a measure of the local vortex strength. A wing can only produce lift when the

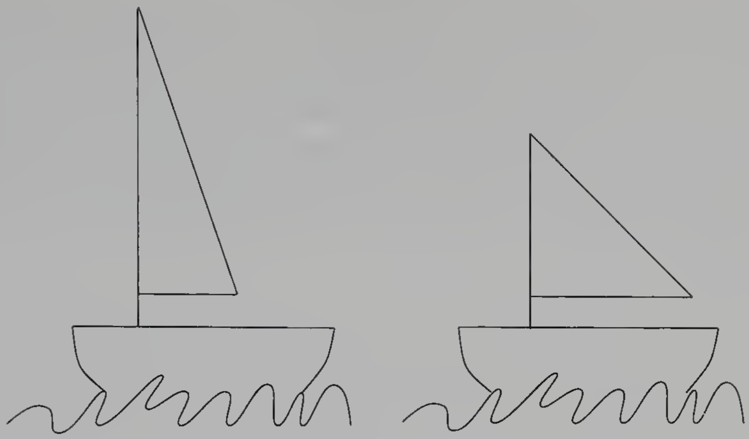

Fig. 9.12 Two boats carrying sails with very different aspect ratios.

circulation along the curve C is nonzero. The above reasoning implies that wingtip vortices are unavoidably associated with the lift produced by an airfoil.

Problem c Consider the wingtip vortex shown in Figure 9.11. You obtained the sign of the circulation $\oint_C \mathbf{v} \cdot d\mathbf{r}$ in problem a. Does this imply that the wingtip vortex rotates in the clockwise direction A of Figure 9.11 or in the counter-clockwise direction B? Use equation (9.22) in your argumentation. You may assume that the vorticity is mostly concentrated at the trailing edge of the wingtips, see Figure 9.11.

The wingtip vortex carries kinetic energy. Since this energy is drawn from the moving aircraft, it is associated with the drag on the aircraft, this is called *induced drag* [55]. Modern aircraft such as the Boeing 747–400, the MD-11, and the Gulfstream 5 have wingtips that are turned upwards. These "winglets" modify the vorticity at the wingtip in such a way that the induced drag on the aircraft is reduced.

Just like aircraft, sailing boats suffer from an energy loss due to a vortex that is generated at the upper part of the sail, see the discussion of Marchaj [67]. (A sail can be considered to be a "vertical wing.") Consider the two boats shown in Figure 9.12 that have sails with the same surface area but with different aspect ratios. The boat on the left will, in general, sail faster. The reason for this is that for the two boats the difference in the wind speed between the two sides of the sail will be roughly identical. This means that for the boat on the right the circulation $\oint_C \mathbf{v} \cdot d\mathbf{r}$ will be larger than that for the boat on the left, simply because the integration contour C is longer. The vorticity generated at the top of the sail of the boat on the right is therefore larger than for the boat on the left. Since this resulting "sailtip" vortex leads to a dissipation of energy, the sail of the boat on the left has a higher efficiency. For the same reason, planes that have to fly with a minimal energy loss (such as gliders) have thin and long wings. In contrast to this, planes that may waste energy in order to fly at a very high speed (such as the Concorde) have wings of a very different shape. Birds follow the same rule: birds that fly relatively slowly but that can glide efficiently over long distances (such as the albatross) have long and thin wings, whereas birds that do not need to fly efficiently (such as a crow) have shorter and thicker wings.

10

Laplacian

The Laplacian of a function consists of a special combination of the second partial derivatives of that function. Before we introduce this quantity, the relation between the second derivative and the curvature of a function is established in Section 10.1. The Laplacian is introduced in Section 10.3 using the physical example of a soap film that minimizes its surface area. The analysis used for this example is introduced in Section 10.2 where a proof is given that the shortest distance between two points is a straight line. The concept of the Laplacian is used in Section 10.5 to study the stability of matter, while in Section 10.6 the implications for the initiation of lightning is considered. Finally, the Laplacian in cylindrical and spherical coordinates is derived, and this is used in Section 10.8 to derive an averaging integral for harmonic functions.

10.1 Curvature of a function

Let us consider a function $f(x)$ that has an extremum. We are free to choose the origin of the x-axis, and the origin is chosen here at the location of the extremum. This means that the function $f(x)$ is stationary at the location $x = 0$, see Figure 10.1. The behavior of the function near its maximum can be studied using the Taylor series (3.11)

$$f(x) = f(0) + \frac{df}{dx}(x = 0)\, x + \frac{1}{2}\frac{d^2 f}{dx^2}(x = 0)\, x^2 + \cdots . \tag{3.11}$$

The point $x = 0$ is a maximum or a minimum which implies by definition that the first derivative vanishes at $x = 0$, so that the function behaves near the extremum as

$$f(x) = f(0) + \frac{1}{2}\frac{d^2 f}{dx^2}(x = 0)\, x^2 + \cdots . \tag{10.1}$$

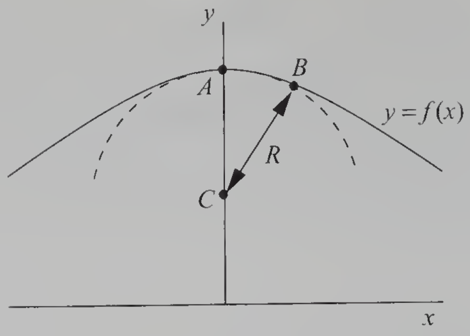

Fig. 10.1 Definition of the geometric variables in the computation of the radius of curvature of a function.

When the first derivative vanishes, the function can have either a maximum or a minimum. In more dimensions, the function can have a minimum in one direction and a maximum in another direction. For this reason the terminology "extremum" is not very accurate. Instead, one refers to a point at which the first derivative vanishes as a *stationary* point. This term is chosen because at such a point, the function according to (10.1) does not vary to first order with the independent variable. In more than one dimension a stationary point is defined by the requirement that the first partial derivatives with respect to *all* variables vanish.

The property of stationary points that the first derivative vanishes corresponds with the fact that at an extremum the slope of the function vanishes. The dominant behavior of the function near its stationary point is given by its *curvature* that is described by the x^2 term in (10.1). In order to characterize this curvature one can define a circle as shown in Figure 10.1 that touches the function $f(x)$ near its extremum. This *tangent-circle* is shown by a dashed line in Figure 10.1. The radius R of this circle measures the curvature of the function at its extremum. For this reason, R is called the *radius of curvature*.

Problem a Convince yourself that a small radius of curvature R implies a large curvature of $f(x)$ and a large radius R corresponds to a small curvature of $f(x)$. What is the shape of $f(x)$ when the radius of curvature tends to infinity $(R \rightarrow \infty)$? Hint: draw this situation.

The radius of curvature can be determined using the points A, B, and C in Figure 10.1.

Problem b Use Figure 10.1 to show that these points have the following coordinates in the (x, y)-plane

$$\mathbf{r}_A = \begin{pmatrix} 0 \\ f(0) \end{pmatrix}, \quad \mathbf{r}_B = \begin{pmatrix} x \\ f(x) \end{pmatrix}, \quad \mathbf{r}_C = \begin{pmatrix} 0 \\ f(0) - R \end{pmatrix}. \tag{10.2}$$

Problem c Since A and B are located on the same circle with C as center, the distance AC equals the distance BC. Use this condition, the coordinates (10.2), and the Taylor expansion (10.1) to derive the following expression

$$x^2 + \left(\frac{1}{2}f''x^2 + R\right)^2 = R^2, \tag{10.3}$$

where f'' denotes the second derivative the extremum: $f'' = d^2 f/dx^2$.

By expanding the square one can obtain an expression for the radius of curvature R. However, when expanding the square, it is only necessary to account for terms up to order x^2. The reason for this is that in the Taylor expansion (10.1) the third order term x^3 as well as all the higher terms have been neglected. (This is because the tangent-circle in Figure 10.1 approximates the function $f(x)$ well only near its maximum.) It is therefore not consistent to retain terms of third and higher order in the remainder of the calculation.

Problem d Use this to derive from (10.3) the following relation between the radius of curvature and the second derivative

$$R = \frac{-1}{f''}. \tag{10.4}$$

This expression relates the radius of the tangent-circle to the second derivative. The approximation of a function by its tangent-circle plays an important role in reflection seismology where it is used to derive the *15-degrees approximation* that accounts for near-vertical wave propagation in the Earth very efficiently [28, 123].

Of course, a stationary point can be either a minimum or a maximum: these two situations are shown in Figure 10.2.

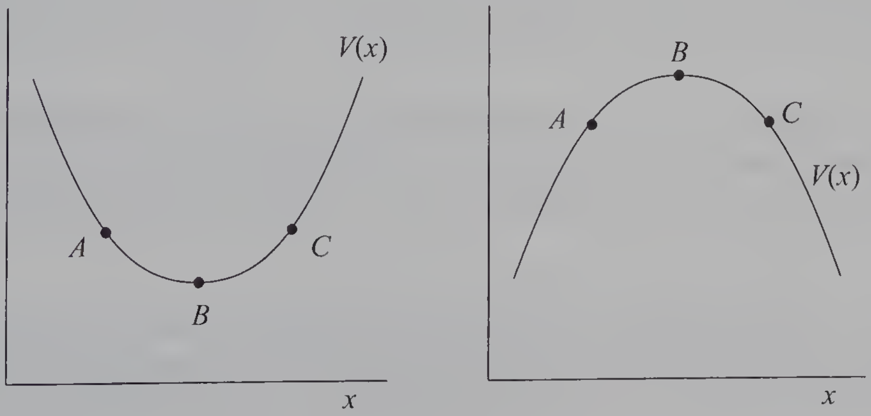

Fig. 10.2 The behavior of a function near a minimum and a maximum.

Problem e Use (10.4) to show that for the curve in the left-hand panel (a minimum) the radius of curvature is negative and that for the curve in the right-hand panel (a maximum) the radius of curvature is positive.

The result in the last problem of course depends critically on the following property of a stationary point:

$$\left.\begin{aligned} &\text{when } f(x) \text{ is a minimum: } \frac{\partial^2 f}{\partial x^2} > 0; \\ &\text{when } f(x) \text{ is a maximum: } \frac{\partial^2 f}{\partial x^2} < 0. \end{aligned}\right\} \tag{10.5}$$

The curve in the left-hand panel of Figure 10.2 is called *concave*, while the curve in the right-hand panel is called *convex*. This implies that a concave curve has a negative radius of curvature and a convex curve has a positive radius of curvature.

The distinction between these different kinds of extrema is crucial to the stability properties of physical systems. Suppose that the function denotes a potential $V(x)$, then according to (5.31) the force associated with potential is given by

$$\mathbf{F}(\mathbf{x}) = -\nabla V(\mathbf{x}). \tag{5.31}$$

In one dimension this expression is given by

$$F(x) = -\frac{dV}{dx}. \tag{10.6}$$

Problem f Show that at the points B in the panels of Figure 10.2 the force vanishes.

Suppose a particle that is influenced by the potential is at rest at one of the points B in Figure 10.2. Since the force vanishes at these points the particle will remain forever at that point. For this reason the points where $\nabla V = 0$ are called the *equilibrium points*.

However, when we move the particle slightly away from the equilibrium points, the force may either push it back towards the equilibrium point, or push it further away. In the first case the equilibrium is *stable*, in the second case it is *unstable*.

Problem g Draw the direction of the force at the points A and C in the two panels of Figure 10.2 and deduce that the equilibrium in the left-hand panel is stable while the equilibrium in the right-hand panel is unstable. Use (10.5) to show that:

$$\left.\begin{aligned} &\text{the equilibrium is stable when } \frac{\partial^2 V}{\partial x^2} > 0; \\ &\text{the equilibrium is unstable when } \frac{\partial^2 V}{\partial x^2} < 0. \end{aligned}\right\} \qquad (10.7)$$

If you find these arguments difficult, you can think of a ball that can move along the curves in the panels of Figure 10.2 in a gravitational field. At point B the ball is in an area that is flat. If it does not move it will remain at that point forever. When the ball in the left-hand panel moves away from the equilibrium point, it rolls uphill and the gravitational force will send it back towards the equilibrium point B. However, when the ball in the right-hand panel moves away from point B, it will roll downhill further from the equilibrium point. The equilibrium at point B in the right-hand panel is unstable. These properties are all related to the second derivative of a function.

10.2 Shortest distance between two points

In this section we give a proof that the shortest distance between two points is a straight line. This may appear to be a trivial problem, but it is shown here because it sets the stage for the Laplacian. It also gives a brief introduction to *variational calculus*. This topic is treated in more detail in Chapter 25.

As an example we show that the shortest curve between two points is a straight line. The problem is shown geometrically in Figure 10.3. Two points A and B are given in the (x, y)-plane; these points are fixed. We are looking for the function $y = h(x)$ that describes the curve with the smallest length.

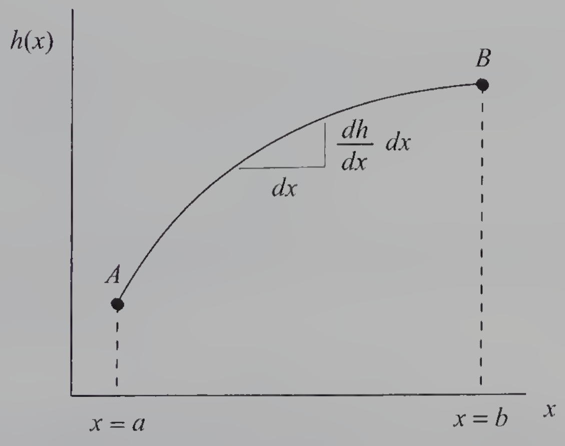

Fig. 10.3 The relation between the derivative of a function and the arclength of the corresponding curve.

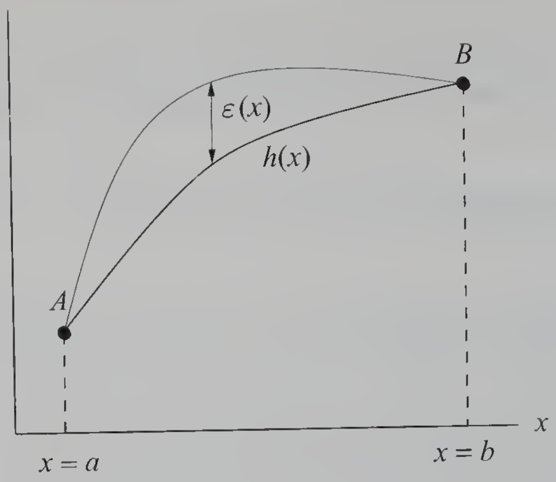

Fig. 10.4 The unperturbed function $h(x)$ and the perturbation $\varepsilon(x)$ that vanishes at the endpoint of the interval.

Problem a We first need to determine the length of the curve given a certain shape $h(x)$. Consider an increment dx of the x-variable. Use the first-order Taylor expansion (3.18) to show that this increment corresponds with an increment $dy = (dh/dx)\,dx$ of the y-variable. Use this result to derive that the length L of the curve is given by:

$$L[h] = \int_a^b \sqrt{1 + h_x^2}\,dx, \tag{10.8}$$

where $h_x = dh/dx$.

Note that the length of the curve depends on the shape $h(x)$ of the curve; for this reason the notation $L[h]$ is used.

We want to find the function $h(x)$ that minimizes the length L. Unfortunately we cannot simply differentiate L with respect to h because $h(x)$ is a function rather than a simple variable. However, we can use the concept of stationarity that was introduced in Section 10.1. When $L[h]$ is minimized, the length of the curve does not change to first order when $h(x)$ is perturbed. Consider Figure 10.4 in which a perturbation $\varepsilon(x)$ is added to the original function $h(x)$. Since the endpoints of the curve are fixed, the perturbation is required to vanish at the endpoints:

$$\varepsilon(a) = \varepsilon(b) = 0. \tag{10.9}$$

In order to solve the problem we need to find the change δL in the length of the curve that is caused by the perturbation $\varepsilon(x)$.

Problem b Replace h in (10.8) by $h + \varepsilon$, carry out a first order Taylor expansion of the integrand with respect to ε and use this to show that the perturbation of the length of the curve is given by

$$\delta L[h] = \int_a^b \frac{h_x \varepsilon_x}{\sqrt{1 + h_x^2}}\, dx. \tag{10.10}$$

In order that the analysis is as transparent as possible we will make an approximation to (10.10) by restricting ourselves to the special case for which the slope of the curve is small. Since $h_x = dh/dx$ denotes the slope of the curve, this corresponds to the condition $h_x \ll 1$. In this case the term h_x in the denominator can be ignored, and the perturbation is given by

$$\delta L[h] = \int_a^b h_x \varepsilon_x dx. \tag{10.11}$$

We show in Section 25.3 that the results of this Section 10.2 also hold when the approximation $h_x \ll 1$ is not used; the only reason for making the approximation is that we do not want the principle of variational calculus to be hidden by analytical complexity.

Expression (10.11) gives the first order perturbation of the length with respect to $\varepsilon(x)$. The condition of stationarity tells us that for the shortest curve, (10.11) must be equal to zero for *any* small perturbation $\varepsilon(x)$. However, we have not yet used the constraint (10.9) which states that the endpoints of the curve are fixed.

Problem c Carry out an integration by parts of (10.11) and use the constraint (10.9) to show that the first order perturbation is given by

$$\delta L[h] = -\int_a^b \frac{d^2 h}{dx^2}\, \varepsilon\, dx. \tag{10.12}$$

For the function $h(x)$ that minimizes the length of the curve this integral must vanish for *all* perturbations $\varepsilon(x)$. This is the case when

$$\frac{d^2 h}{dx^2} = 0. \tag{10.13}$$

This expression states that the curve that has the smallest length has no curvature, and it therefore a straight line.

Problem d Let the y-coordinates of the points A and B be denoted by y_A and y_B respectively. Integrate (10.13) subject to the condition that the curve goes through the points A and B and convince yourself that the solution is indeed given by a straight line.

In the following section this minimization problem is generalized to two dimensions.

10.3 Shape of a soap film

In the last section you derived that the requirement that a curve between two points has the smallest length implies that its curvature vanishes. Suppose we consider a two-dimensional surface in three dimensions, and that we seek the surface that minimizes its surface area given the locations of the edges of the surface. Physically this problem is described by a soap film that is suspended within a wire frame. The surface tension of the soap film minimizes its surface area subject to the constraint that the edges of the soap film are fixed. In this section, the shape of the soap film is described by giving the z-coordinate as a function of the x- and y-coordinates: $z = h(x, y)$.

At first sight one might guess that the soap film will be plane, because any deviations of the soap film from this plane will increase its surface area. Since the curvature of a plane vanishes in the x-direction as well as in the y-direction, the analogy with expression (10.13) suggests that the soap film satisfies the following equations: $\partial^2 h/\partial x^2 = \partial^2 h/\partial y^2 = 0$. That this is not in general true can be seen in the soap film shown in Figure 10.5. In this example the wire frame that defines the

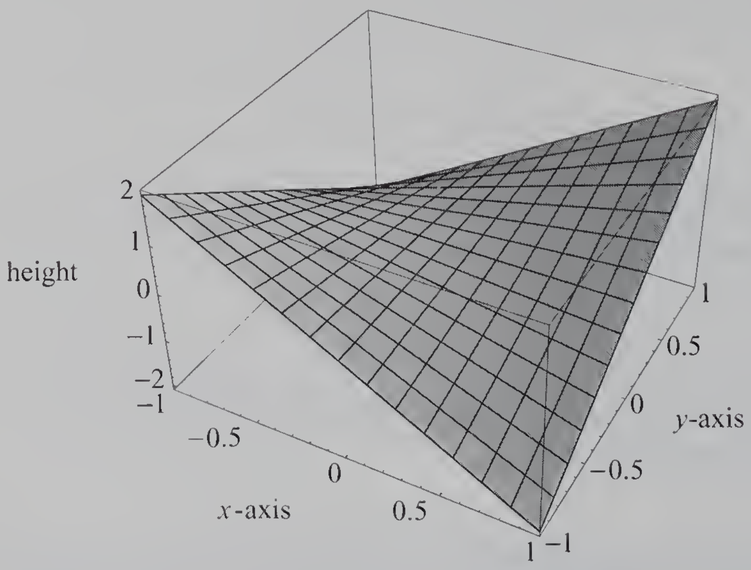

Fig. 10.5 The shape of a soap film whose edges are fixed at the outer edges of the box.

edge of the soap film is not confined to a plane. This results in a soap film that is curved and hence it is not possible that both $\partial^2 h/\partial x^2$ and $\partial^2 h/\partial y^2$ are equal to zero. In this section we derive an equation for the shape of the soap film.

The position of a point on the soap film is given by the following position vector:

$$\mathbf{r} = \begin{pmatrix} x \\ y \\ h(x, y) \end{pmatrix}. \tag{10.14}$$

In problem e of Section 4.4 you showed that the surface area dS of a sphere that corresponds to increments $d\theta$ and $d\varphi$ of the angles on the sphere is given by

$$dS = \left| \frac{\partial \mathbf{r}}{\partial \theta} \times \frac{\partial \mathbf{r}}{\partial \varphi} \right| d\theta d\varphi. \tag{4.34}$$

Problem a Use the reasoning used in problem e of Section 4.4 to show that increments of the surface area of the soap film satisfy the following expression

$$dS = \left| \frac{\partial \mathbf{r}}{\partial x} \times \frac{\partial \mathbf{r}}{\partial y} \right| dxdy. \tag{10.15}$$

Problem b Use (10.14) and (10.15) to derive that the total surface area is given by

$$S = \iint \sqrt{1 + |\nabla h|^2}\, dxdy. \tag{10.16}$$

Note the analogy between this expression and (10.8).

Since the surface tension that governs the shape of the soap film tends to minimize the surface area of the soap film, the shape $h(x, y)$ follows from the requirement that $h(x, y)$ is the function that minimizes the surface area (10.16). As in Section 10.2 the solution follows from the requirement that for the function $h(x, y)$ that minimizes the surface area, the surface area is stationary for perturbations of $h(x, y)$. This means that when $h(x, y)$ is replaced by $h(x, y) + \varepsilon(x, y)$ the first order change of the surface area S vanishes.

Problem c Make the substitution $h(x, y) \rightarrow h(x, y) + \varepsilon(x, y)$, use the identity $|\nabla h|^2 = (\nabla h \cdot \nabla h)$ in (10.16) and linearize the result in $\varepsilon(x, y)$ to derive that the perturbation of the surface area is to first order in $\varepsilon(x, y)$ given by

$$\delta S[h] = \iint \frac{(\nabla h \cdot \nabla \varepsilon)}{\sqrt{1 + |\nabla h|^2}} dxdy. \tag{10.17}$$

In order to concentrate on the essentials we assume that the deflection of the soap surface is small, which means that $|\nabla h| \ll 1$. Under this assumption the $|\nabla h|$ term in the denominator can be ignored and the perturbation is given by

$$\delta S[h] = \iint (\nabla h \cdot \nabla \varepsilon)\, dxdy. \tag{10.18}$$

In the problem we are considering here, the edge of the soap film is kept at a fixed location. This means that the perturbation $\varepsilon(x, y)$ must vanish at the edge of the soap film:

$$\varepsilon(x, y) = 0 \qquad \text{at the edge of the soap film.} \tag{10.19}$$

This constraint has not yet been taken into account; it is incorporated in the following two problems.

Problem d Derive the identity $(\nabla h \cdot \nabla \varepsilon) = \nabla \cdot (\varepsilon \nabla h) - \varepsilon \nabla \cdot \nabla h$.

Problem e Insert this result in (10.18), apply Gauss's theorem to the first term, and use (10.19) to show that the resulting integral over the edge of the surface vanishes, so that the perturbation of the surface area is given by

$$\delta S[h] = -\iint \varepsilon\, (\nabla \cdot \nabla h)\, dxdy. \tag{10.20}$$

The second derivative $(\nabla \cdot \nabla h)$ is called the *Laplacian* of h. This operator is often denoted by the notation Δ. However, since the term $\nabla \cdot \nabla$ is reminiscent of the square of the vector ∇, the notation ∇^2 is also used, and in this book this latter notation is the one we will mostly use for the Laplacian. The requirement that the first order perturbation of the soap film vanishes implies that $(\nabla \cdot \nabla h)$ must be equal to zero. The soap film therefore satisfies the following differential equation:

$$\nabla^2 h = 0. \tag{10.21}$$

In mathematical physics this equation is called the *Laplace equation.* Before analyzing this equation in more detail we first define the Laplacian more precisely.

Problem f Use the definition $\nabla^2 = \nabla \cdot \nabla$ to show that

$$\nabla^2 = \text{div grad} = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} \quad \text{(in two dimensions)} . \tag{10.22}$$

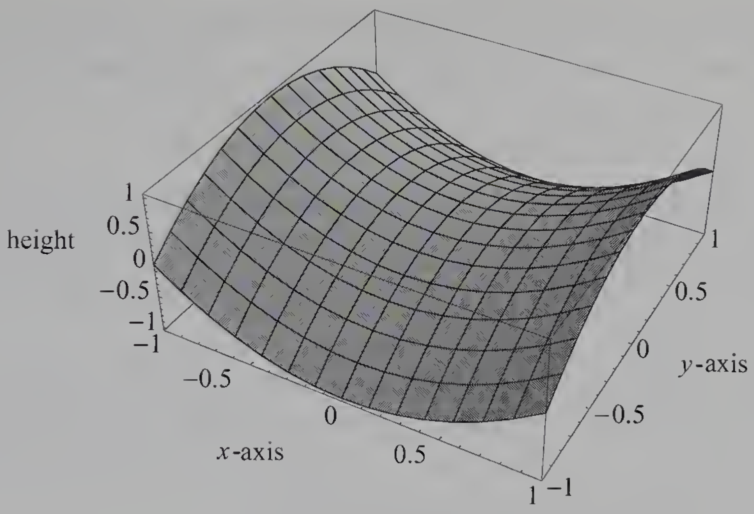

Fig. 10.6 The shape of the same soap film as in the previous figure after a rotation through 45 degrees. (New edges define the area of this soap film.) The shape of the soap film is given by the function $h(x, y) = x^2 - y^2$; the shape of the soap film in the previous figure is given by $h(x, y) = 2xy$.

Analogously the Laplacian in three dimensions is defined as

$$\nabla^2 = \text{div grad} = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2} \qquad \text{(in three dimensions)} . \quad (10.23)$$

Let us now consider the shape of the soap film that is shown in Figure 10.5. At the beginning of this section it was argued that the curvatures $\partial^2 h/\partial x^2$ and $\partial^2 h/\partial y^2$ of the soap film in the x- and y-directions could not both be equal to zero. Instead, (10.21) implies that the *sum* of the curvatures in the x- and y-directions vanishes. This can be seen by rotating the soap film through 45 degrees as shown in Figure 10.6.

Problem g Sketch cross sections of the surface in Figure 10.6 along the x- and y-axes, and show that for that surface the curvature in the x-direction is positive and the curvature in the y-direction is negative.

Problem h The surface in Figure 10.6 is given by

$$h(x, y) = x^2 - y^2. \qquad (10.24)$$

Show that for this surface $\nabla^2 h = 0$ and that the curvatures in the x-direction and in the y-direction cancel.

These results imply that the soap film behaves in a fundamentally different way than the rubber band between two fixed points that we treated in the previous section. The rubber band follows a straight line and the curvature is equal to zero while for the soap film the *sum* of the curvatures in orthogonal directions vanishes. However, as shown in Figures 10.5 and 10.6 this does not imply that the soap film is a planar surface. The fact that the curvature terms $\partial^2 h/\partial x^2$ and $\partial^2 h/\partial y^2$ have opposite signs implies that when the function is concave in one direction it must be convex in the other direction. This means that when the function is a maximum in one direction, it must be a minimum in the other direction. The surface therefore has the shape of saddle.

Let us now consider a stationary point of the function: that is a point where $\partial h/\partial x = \partial h/\partial y = 0$. This point can be neither a minimum nor a maximum because when it is a minimum for variations in one direction, it *must* be a maximum for variations in the other direction. This means that:

Theorem *A function that satisfies $\nabla^2 h = 0$ cannot have an extremum; the function can only have a maximum or minimum at the edge of the domain on which it is defined.*

We will see in Section 10.5 that this has important consequences. When a function is a minimum all the second derivatives must analogously to (10.5) be positive, while for an extremum that is a maximum all the second derivatives must be negative.

Problem i Show that:

$$\left.\begin{array}{l} \text{when } h(\mathbf{r}) \text{ is a minimum: } \nabla^2 h > 0; \\ \\ \text{when } h(\mathbf{r}) \text{ is a maximum: } \nabla^2 h < 0. \end{array}\right\} \qquad (10.25)$$

Problem j The above condition is a necessary but not a sufficient condition. (This means that every minimum must satisfy the requirement $\nabla^2 h > 0$, but conversely when $\nabla^2 h > 0$ is satisfied the function is not necessarily a minimum.) Show this by analyzing the stationary point of the function $h(x, y, z) = x^2 + y^2 - z^2$.

10.4 Sources of curvature

The main point of the previous section is that a function that satisfies the Laplace equation (10.21) can have nonzero curvature. In general there are two reasons why a soap film can be curved. The first reason is that the edges of the soap film are not confined to a common plane, as shown in Figures 10.5 and 10.6. Since the edges are part of the soap film, this necessarily implies that the soap film cannot be confined to a plane.

There is another reason why a soap film can be curved. Let us consider a soap film whose edges are located in the (x, y)-plane:

$$h(x, y) = 0 \qquad \text{at the edge.} \tag{10.26}$$

However, let us suppose that gravity pulls the soap film down. Such a soap film will bend down from the edges and will therefore be curved. This means that an external force can also lead to curvature of the soap film.

In this section we derive the differential equation for a soap film that is subject to a gravitational force. This is achieved by minimizing the total energy of the soap film. The gravitational energy is given by $\rho g h$, where ρ is the mass-density per unit area and g is the acceleration of gravity. The surface energy of the soap film is some positive constant k times the surface area of the soap film.

Problem a Use these results and (10.16) to show that the total energy is given by:

$$E = \iint \left(k\sqrt{1 + |\nabla h|^2} + \rho g h \right) \, dxdy. \tag{10.27}$$

The shape of the soap film is determined by the condition that the energy is a minimum.

Problem b In the previous section we minimized the first term in the integrand. Generalize the derivation of the previous section to take the second term in (10.27) into account as well to derive that the soap film under gravity satisfies the following differential equation:

$$\nabla^2 h = \frac{\rho g}{k}. \tag{10.28}$$

This equation is called *Poisson's equation*. The gravitational force has the effect that the total curvature of the soap film can be nonzero. This corresponds to the reasoning at the beginning of this section that gravity will make the soap film sag, so that the total curvature can be nonzero despite the fact that the edges of the soap film are confined to a plane. Note that gravity acts as a source term in the differential equation (10.28).

There is an interesting analogy between the soap film in a gravitational field and the electric potential that is generated by electric charges. According to equation (6.23) the electric field generated by a charge density ρ is given by

$$(\nabla \cdot \mathbf{E}) = \rho(\mathbf{r})/\varepsilon_0. \tag{6.23}$$

By analogy with (5.31) the potential associated with this electric field is given by

$$\mathbf{E} = -\nabla V. \tag{10.29}$$

Problem c Show that the potential satisfies

$$\nabla^2 V = -\rho/\varepsilon_0. \tag{10.30}$$

This means that the electric potential satisfies Poisson's equation (10.28) as well. The electric charge acts as the source of the electric potential just as the gravitational force acts as the source of the deflection of the soap film.

The analogy between the deflection of the soap film and the electric potential is interesting. Let us consider once more a soap film that is not subject to a gravitational force. The equation (10.28) of the soap film followed from the requirement that the squared-gradient $|\nabla h|^2$ integrated over the surface was minimized, because then the term $\sqrt{1+|\nabla h|^2}$ in (10.16) is minimized as well. The analogy with the electric potential means that in free space ($\rho = 0$) the electric potential behaves in such a way that the squared-gradient $|\nabla V|^2$of the potential integrated over the volume is minimized. However, according to (10.29) the gradient of the electric potential is the electric field. This means that the electric field behaves in such a way that the volume integral $\int |\mathbf{E}|^2 \, d^3x$ is minimized. However, this quantity is nothing but the energy of a static electric field [53]. This implies that the electric field is distributed in such a way that the energy of the electric field is minimized.

10.5 Instability of matter

The title of this section may surprise you, but the results that will be obtained here imply that according to the laws of classical physics the structure of matter and the mass distribution in the universe cannot be in a stable equilibrium. This result was derived by Earnshaw in 1842 in his work "On the nature of the molecular forces which regulate the constitution of the luminiferous ether" [36]. Let us consider a particle in three dimensions that is subject to a potential $V(\mathbf{r})$ that accounts for the gravitational attraction by other masses and the electrostatic force due to other charges. Let us suppose that at some point in space the particle is in equilibrium. This means that the force acting on the particle vanishes at that point, or equivalently that the gradient of the potential vanishes at that point: $\nabla V = 0$.

Problem a Generalize expression (10.7) to three dimensions to show that this equilibrium point is only stable when the curvature of the potential in the three coordinate directions is positive:

$$\frac{\partial^2 V}{\partial x^2} > 0 \quad \text{and} \quad \frac{\partial^2 V}{\partial y^2} > 0 \quad \text{and} \quad \frac{\partial^2 V}{\partial z^2} > 0. \tag{10.31}$$

According to (10.30), in free space ($\rho = 0$) the potential that corresponds to the gravitational force exerted by other masses, and the electrostatic force generated by other charges, satisfies Laplace's equation $\nabla^2 V = 0$.

Problem b Show that Laplace's equation implies that when the curvature of the potential is positive in one direction, the curvature must be negative in at least one other direction.

Problem c Use this to deduce that when the equilibrium point is stable to perturbations in one direction, it must be unstable for perturbations in at least one other direction.

An equilibrium point is in general unstable when it is unstable for perturbations in at least one of the directions. Using problem c this means that an equilibrium point for the potential $V(\mathbf{r})$ that satisfies Poisson's equation cannot be stable.

This has far-reaching consequences. Let us consider a crystal. Within the framework of classical physics, each ion in the crystal moves in an electric potential that is generated by all the other ions in the crystal. This potential satisfies Laplace's equation at the location of the ion that we are considering because the net charge density ρ of the *other* ions is zero at that point. This means that the motion of the ion at its equilibrium point is not stable. When this is the case, the crystal is not stable because small perturbations of each ion from its equilibrium point lead to unstable motions of the ions. This implies that according to the laws of classical physics, matter is not stable! This result, known as *Earnshaw's theorem* [36], states that the equilibrium points of any configuration of static electrical, magnetic, and gravitational fields are unstable.

Hopefully you are convinced that matter is stable, whatever the unsettling prediction of Earnshaw's theorem may be. The only way that we can resolve this paradox is if matter does not satisfy the equations of classical mechanics and electrostatics. This implies that quantum effects must play a crucial role in the stability of matter.

Earnshaw's theorem also applies to the gravitational field and it states that equilibrium points in the gravitational field are not stable. This means that a universe in equilibrium would not be stable! Let us first consider our solar system. It is essential for the solar system that the planets and their moons are in motion. This means that the solar system is not in a static equilibrium. In fact, the motion of the planets and moons is governed by the combination of the gravitational force plus the inertia force that is associated with the motion of the planets. The miracle of planetary motion is that the gravitational *motion* is stable. (You showed in Section 6.5 that the motion of planetary orbits is only stable in less than four spatial dimensions.) The situation is comparable to the movement of a bicycle which derives its stability from the motion of the wheels that rotate. In the same way, it is the combination of the gravitational field and the inertial forces due to the motion of the planets that leads to stable planetary orbits. On scales much larger than the solar system, relativistic effects are important and Poisson's equation is not sufficient to describe the gravitational field on a cosmological scale [76].

In 1842, Earnshaw [36] carried out his work not to study the structure of matter, but to study the structure of the ether, which at that time was seen as the carrier of electromagnetic waves. He viewed the ether as a system of interacting particles and showed that the interaction between these particles could not be governed by Newton's law of gravitation using the analysis shown in this section. From the requirement of stability he derived conditions for the potential that governs the interaction between etheral particles. Note that in 1842 Earnshaw could not invoke quantum mechanics in his description of interacting microscopic particles because that theory had not yet been formulated.

10.6 Where does lightning start?

During a thunderstorm, the vertical motion of ice particles causes the separation of positive and negative electric charges in the atmosphere. This charge separation leads to an electric field within the atmosphere. When the field strength at a certain location is sufficiently strong, atoms are ionized and a current flows. This current induces further ionization which extends the path of the current. This physical process describes the events that initiate a lightning bolt. In principle there seems to be no reason why a lightning bolt cannot start in mid-air, far away from the Earth's surface and far from the electrical charges that induce the electric field. In this section we show that lightning can only start at specific locations.

To see this we study the Laplacian of the square of the length of a vector field **E**.

Problem a Compute the Laplacian of $E^2 = E_x^2 + E_y^2 + E_z^2$ and show that it is given by

$$\nabla^2 E^2 = 2\left(|\nabla E_x|^2 + \left|\nabla E_y\right|^2 + |\nabla E_z|^2\right) + 2\left(\mathbf{E} \cdot \nabla^2 \mathbf{E}\right). \tag{10.32}$$

(Up to this point the Laplacian has always acted on a scalar; when it acts on a vector as in (10.32) it means that the Laplacian of every component is taken so that $\nabla^2\mathbf{E}$ stands for a vector with components $\nabla^2 E_x$, $\nabla^2 E_y$ and $\nabla^2 E_z$, respectively.)

Let us now consider a vector field $\mathbf{E}$ that satisfies the Laplace equation:

$$\nabla^2 \mathbf{E} = 0. \tag{10.33}$$

Problem b Show that a vector field that satisfies the Laplace equation (10.33) satisfies the following inequality:

$$\nabla^2 E^2 \geq 0. \tag{10.34}$$

Problem c Use this to show that the strength of the vector field (E^2) cannot have a maximum in the region where (10.33) is satisfied.

Note that (10.34) does not exclude the possibility that E^2 has a minimum.

The result you derived in problem c is all we need to solve our problem. Away from the charges in the atmosphere, a static electric field satisfies (10.33), see reference [53]. According to problem c, the electric field strength E^2 cannot have a maximum in the region of the atmosphere where there are no charges that are the source of the field. Lightning will initiate where the field strength is strongest. This means that lightning must initiate either in the regions where Laplace equation (10.33) does not hold (at the charges that generate the electric field), or at the boundary of the area (the Earth's surface). This implies that lighting must start either at the charges that generate the field or at the Earth's surface.

Problem d The source of the magnetic field of the Earth is located in the Earth's core. Away from this source, the magnetic field satisfies Laplace equation (10.33) (at least when the field is stationary)[53]. Show that the magnetic field of the Earth cannot have its maximum strength outside the Earth.

10.7 Laplacian in spherical and cylindrical coordinates

In many applications it is useful to use the Laplacian in spherical or cylindrical coordinates. In principle this result can be derived by applying the transformation rules to the second partial derivatives in the Laplacian when changing from Cartesian to spherical or cylindrical coordinates. However, this route is unnecessarily complex, especially since we have done most of the work already. The key element in the derivation was derived in problem f of Section 10.3 where you showed that the Laplacian is the divergence of the gradient: $\nabla^2 = \text{div}\,\text{grad}$. In Sections 5.6 and 6.4 you have already derived expressions for the gradient and the divergence in spherical and cylindrical coordinates, and all you need to do is to insert the expression for the gradient in curvilinear coordinates into the expression of the divergence in curvilinear coordinates.

Problem a Use (5.48) and (6.31) to show that the Laplacian of a function f in cylindrical coordinates is given by

$$\nabla^2 f = \frac{1}{r}\frac{\partial}{\partial r}\left(r\frac{\partial f}{\partial r}\right) + \frac{1}{r^2}\frac{\partial^2 f}{\partial \varphi^2} + \frac{\partial^2 f}{\partial z^2}. \tag{10.35}$$

Problem b Find the expressions in Sections 5.6 and 6.4 that allow you to derive that the Laplacian of a function f in spherical coordinates is given by

$$\nabla^2 f = \frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2\frac{\partial f}{\partial r}\right) + \frac{1}{r^2 \sin\theta}\frac{\partial}{\partial \theta}\left(\sin\theta\frac{\partial f}{\partial \theta}\right) + \frac{1}{r^2 \sin^2\theta}\frac{\partial^2 f}{\partial \varphi^2}. \tag{10.36}$$

In later parts of this book extensive use is made of these expressions of the Laplacian.

10.8 Averaging integrals for harmonic functions

In this section we focus on functions f that satisfy Laplace's equation (10.21), that is $\nabla^2 f = 0$. Functions whose Laplacian is zero are called *harmonic functions*. They play an important role in mathematical physics. We will show in Section 16.1 that the real and imaginary parts of analytic functions in the complex plane are harmonic functions. Let us first focus on a harmonic function $f(x, y)$ in two dimensions. In this section we derive that the function value at a certain point is equal to the average of that function over a circle centered around that point with an *arbitrary* radius. To see this we use a system of cylindrical coordinates and choose the origin of the cylindrical coordinates at the point that we consider. (Remember that we are free to choose the origin of the coordinate system.)

Problem a Use expression (10.35) to show that f satisfies the following differential equation:

$$\frac{1}{r}\frac{\partial}{\partial r}\left(r\frac{\partial f}{\partial r}\right) + \frac{1}{r^2}\frac{\partial^2 f}{\partial \varphi^2} = 0. \tag{10.37}$$

Problem b Integrate this expression over a disk of radius R centered at the origin and derive that

$$\int_0^R \int_0^{2\pi} \left[\frac{\partial}{\partial r}\left(r\frac{\partial f}{\partial r}\right) + \frac{1}{r}\frac{\partial^2 f}{\partial \varphi^2}\right] d\varphi dr = 0. \tag{10.38}$$

(Note the powers of r in this expression.)

Problem c Carry out the φ-integration in the last term of the integrand to show that this term gives a vanishing contribution.

It is convenient to introduce at this point the average $\bar{f}(r)$ of the function f over a circle with radius r:

$$\bar{f}(r) \equiv \frac{1}{2\pi} \int_0^{2\pi} f(r, \varphi)\, d\varphi. \tag{10.39}$$

Problem d Use (10.38) and the result of problem c to derive that $\bar{f}$ satisfies the following equation:

$$\left[r \frac{\partial \bar{f}}{\partial r} \right]_{r=0}^{r=R} = 0. \tag{10.40}$$

Problem e This expression holds for any radius R, hence $r \partial \bar{f}/\partial r$ is independent of r, so that it is a constant: $r \partial \bar{f}/\partial r = C$. Integrate this expression to derive that

$$\bar{f}(r) = C \ \ln r + A, \tag{10.41}$$

where A is an unknown integration constant.

Problem f The integration constant C must be equal to zero because $\bar{f}(r)$ is finite as $r \to 0$. Evaluate (10.41) at the origin, determine the constant A and show that $f(r)$ satisfies

$$f(r = 0) = \frac{1}{2\pi} \int_0^{2\pi} f(r, \varphi)\, d\varphi. \tag{10.42}$$

This expression states that the value of a harmonic function at the origin is the average of the function over a circle with arbitrary radius. The amazing property is that this holds for any value of the radius, provided f is harmonic everywhere within the circle with radius r. Note that we are free to choose the origin of the coordinate system so that the averaging integral (10.42) holds for any point.

The averaging integral can also be used to prove that a harmonic function cannot have a minimum or a maximum in the region where it is defined. Suppose the function had a maximum at a certain location. There then exists a circle around this point where the function has smaller values than at the maximum (otherwise it would not be a maximum). Equation (10.42) cannot hold for this circle because the right-hand side would be smaller than the left-hand side. Therefore it is impossible for f to have a maximum.

Problem g Generalize the derivation in this section to spherical coordinates and derive the following expression for harmonic functions in three dimensions

$$f(r=0)=\frac{1}{4\pi}\int_0^{\pi}\int_0^{2\pi} f(r,\theta,\varphi)\,\sin\theta\; d\varphi d\theta . \tag{10.43}$$

Problem h Show that the right-hand side of this expression is the average of that function over a sphere with arbitrary radius r centered around that point $r=0$.

This means that, in three dimensions, the value of a harmonic function at a certain point is equal to the average of the function over a sphere with arbitrary radius r that is centered on that point.

11

Conservation laws

In physics one frequently handles the change of a property with time by considering properties that do *not* change with time. For example, when two particles collide elastically, the momentum and the energy of each particle may change. However, this change can be found from the consideration that the total momentum and energy of the system are conserved. Often in physics, such conservation laws are the main ingredients for describing a system. In this chapter we deal with conservation laws for continuous systems. These are systems in which the physical properties are a continuous function of the space coordinates. Examples are the motion in a fluid or solid, and the temperature distribution in a body. The introduced conservation laws are not only of great importance in physics, they also provide worthwhile exercises in the use of vector calculus introduced in the previous chapters.

11.1 General form of conservation laws

In this section a general derivation of conservation laws is given. Suppose we consider a physical quantity Q. This quantity could denote the mass density of a fluid, the heat content within a solid or any other type of physical variable. In fact, there is no reason why Q should be a scalar, it could also be a vector (such as the momentum density) or a higher order tensor. Let us consider a volume V in space that does not change with time. This volume is bounded by a surface ∂V. The total amount of Q within this volume is given by the integral $\int_V Q dV$. The rate of change of this quantity with time is given by $(\partial/\partial t)\int_V Q dV$.

In general, there are two reasons for the quantity $\int_V Q dV$ to change with time. First, the field Q may have sources or sinks within the volume V. The net source of the field Q per unit volume is denoted by the symbol S. The total source of Q within the volume is simply the volume integral $\int_V S dV$ of the source density. Second, it may be that the quantity Q is transported in the medium. With this transport process is associated a current $\mathbf{J}$ of the quantity Q. As an example one can think of Q as being the mass-density of a fluid. In that case $\int_V Q dV$ is the total mass of the fluid in the volume. This total mass can change because there is a source of fluid within the volume (i.e. a tap or a bathroom sink), or the total mass may change because of the flow through the boundary of the volume.

The rate of change of $\int_V Q dV$ by the current is given by the *inward* flux of the current $\mathbf{J}$ through the surface ∂V. If we retain the convention that the surface element $d\mathbf{S}$ points out of the volume, the inward flux is given by $-\oint_{\partial V} \mathbf{J} \cdot d\mathbf{S}$. Together with the rate of change due to the source density S within the volume this implies that the rate of change of the total amount of Q within the volume satisfies:

$$\frac{\partial}{\partial t} \int_V Q dV = -\oint_{\partial V} \mathbf{J} \cdot d\mathbf{S} + \int_V S dV. \tag{11.1}$$

Using Gauss's law (8.1), the surface integral on the right-hand side can be written as $-\int_V (\nabla \cdot \mathbf{J}) dV$, so that the expression above is equivalent with

$$\frac{\partial}{\partial t} \int_V Q dV + \int_V (\nabla \cdot \mathbf{J}) dV = \int_V S dV. \tag{11.2}$$

Since the volume V is assumed to be fixed with time, the time-derivative of the volume integral is the volume integral of the time-derivative: $(\partial/\partial t) \int_V Q dV = \int_V (\partial Q/\partial t) dV$. It should be noted that (11.2) holds for *any* volume V. If the volume is an infinitesimal volume, the volume integrals in (11.2) can be replaced by the integrand multiplied by the infinitesimal volume. This means that (11.2) is equivalent to:

$$\frac{\partial Q}{\partial t} + (\nabla \cdot \mathbf{J}) = S. \tag{11.3}$$

This is the general form of a conservation law in physics; it simply states that the rate of change of a quantity is due to the sources (or sinks) of that quantity and due to the divergence of the current of that quantity. Of course, the general conservation law (11.3) is not very meaningful as long as we do not provide expressions for the current $\mathbf{J}$ and the source S. In this section we will see some examples in which the current and the source follow from physical theory, but we will also encounter examples in which they follow from an "educated" guess.

Equation (11.3) will not be completely new to you. In Section 8.4 the probability-density current for a quantum-mechanical system was derived.

Problem a Use the derivation in this section to show that expression (8.16) can be written as

$$\frac{\partial}{\partial t} |\psi|^2 + (\nabla \cdot \mathbf{J}) = 0, \tag{11.4}$$

with $\mathbf{J}$ given by (8.18).

This equation constitutes a conservation law for the probability density of a quantum particle. Note that (8.16) could be derived rigorously from the Schrödinger equation (8.14), so that the conservation law (11.4) and the expression for the current **J** follow from the basic equation of the system.

Problem b Why is the source term on the right-hand side of (11.4) equal to zero?

11.2 Continuity equation

In this section we consider the conservation of mass in a continuous medium such as a fluid or a solid. In that case, the quantity Q is the mass-density ρ. If we assume that mass is neither created nor destroyed, the source term vanishes: $S = 0$. The vector **J** is the mass current and denotes the flow of mass per unit volume. Consider a small volume δV. The mass within this volume is equal to $\rho \delta V$. If the velocity of the medium is denoted by $\mathbf{v}$, the mass flow is given by $\rho \delta V \mathbf{v}$. Dividing this by the volume δV one obtains the mass flow per unit volume; this quantity is called the mass-density current:

$$\mathbf{J} = \rho \mathbf{v}. \tag{11.5}$$

Using these results, the principle of the conservation of mass can be expressed as

$$\frac{\partial \rho}{\partial t} + \nabla \cdot (\rho \mathbf{v}) = 0. \tag{11.6}$$

This expression plays an important role in continuum mechanics, and is called the *continuity equation*.

Up to this point the reasoning has been based on a volume V that did not change with time. This means that our treatment was strictly Eulerian: we considered the change of physical properties at a fixed location. As an alternative, a Lagrangian description of the same process can be given. In such an approach one specifies how physical properties change as they are moved along by the flow. In this approach one seeks an expression for the total time-derivative d/dt of physical properties rather than expressions for the partial derivative $\partial/\partial t$. It follows from (5.40) that these two derivatives are related in the following way:

$$\frac{d}{dt} = \frac{\partial}{\partial t} + (\mathbf{v} \cdot \nabla). \tag{11.7}$$

This distinction between the total time-derivative and the partial time-derivative is treated in detail in Section 5.5.

Problem a Show that the total derivative of the mass-density is given by:

$$\frac{d\rho}{dt} + \rho(\nabla \cdot \mathbf{v}) = 0. \tag{11.8}$$

Problem b This expression gives the change in the density when one follows the flow. Let us consider an infinitesimal volume δV that is carried around with the flow. The mass of this volume is given by $\delta m = \rho \delta V$. The volume moves with the flow, therefore there is no mass transport across the boundary of the volume. This means that the mass within that volume is conserved: $\delta \dot{m} = 0$. The dot denotes the total time-derivative, *not* the partial time-derivative. Use this expression and (11.8) to show that $(\nabla \cdot \mathbf{v})$ is the rate of change of the volume normalized by size of the volume:

$$\frac{\delta \dot{V}}{\delta V} = (\nabla \cdot \mathbf{v}). \tag{11.9}$$

We have learned a new meaning of the divergence of the velocity field: it equals the relative change in volume per unit time.

11.3 Conservation of momentum and energy

In the description of a point mass in classical mechanics, the conservation of momentum and energy can be derived from Newton's third law. The same is true for a continuous medium such as a fluid or a solid. In order to formulate Newton's law for a continuous medium we start with a Lagrangian point of view and consider a volume δV *that moves with the flow*. The mass of this volume is given by $\delta m = \rho \delta V$. This mass is constant because the volume is defined to move with the flow, hence mass cannot flow into or out of the volume. Let the force per unit volume be denoted by $\mathbf{F}$, so that the total force acting on the volume is $\mathbf{F}\delta V$. The force $\mathbf{F}$ contains both forces generated by external agents (such as gravity) and internal agents such as the pressure force $-\nabla p$ or the effect of internal stresses $(\nabla \cdot \boldsymbol{\sigma})$. The stress tensor $\boldsymbol{\sigma}$ is treated in Section 22.10. Newton's law applied to the volume δV takes the form:

$$\frac{d}{dt}(\rho \delta V \mathbf{v}) = \mathbf{F}\delta V. \tag{11.10}$$

Since the mass $\delta m = \rho \delta V$ is constant with time it can be taken outside the derivative in (11.10). Dividing the resulting expression by δV leads to the Lagrangian form of the equation of motion:

$$\rho \frac{d\mathbf{v}}{dt} = \mathbf{F}. \tag{11.11}$$

Note that the density appears *outside* the time-derivative, despite the fact that the density may vary with time. Using the prescription (11.7) one obtains the Eulerian form of Newton's law for a continuous medium:

$$\rho \frac{\partial \mathbf{v}}{\partial t} + \rho \mathbf{v} \cdot \nabla \mathbf{v} = \mathbf{F}. \tag{11.12}$$

This equation is not yet in the general form (11.3) of conservation laws because in the first term on the left-hand side we have the density *times* a time-derivative, and because the second term on the left-hand side is not the divergence of some current.

Problem a Use (11.12) and the continuity equation (11.6) to show that:

$$\frac{\partial(\rho\mathbf{v})}{\partial t} + \nabla\cdot(\rho\mathbf{v}\mathbf{v}) = \mathbf{F}. \tag{11.13}$$

This expression does take the form of a conservation law; it expresses that the momentum (density) $\rho\mathbf{v}$ is conserved. (For brevity we will often not include the affix "density" in the description of the different quantities, but remember that all quantities are given per unit volume.) The source of momentum is given by the force $\mathbf{F}$, and this reflects that forces are the cause of changes in momentum. In addition there is a momentum current $\mathbf{J} = \rho\mathbf{v}\mathbf{v}$ that describes the transport of momentum by the flow. This momentum current is not a simple vector, it is a dyad and hence is represented by a 3 × 3 matrix. This is not surprising since the momentum is a vector with three components and each component can be transported in three spatial directions.

Sometimes there is a certain arbitrariness in what we call the current and what we call the source. As an example let us consider (11.13) again. According to (5.12) the pressure force is given by $\mathbf{F} = -\nabla p$. If there are no other forces acting on the fluid, the right-hand side of (11.13) is given by $-\nabla p$. This term does not have the form of the divergence, but we can rewrite it by using that:

$$\frac{\partial p}{\partial x_i} = \sum_j \frac{\partial}{\partial x_j}\left(p\delta_{ij}\right), \tag{11.14}$$

where δ_{ij} is the Kronecker delta. This quantity is defined as follows:

$$\delta_{ij} = \begin{cases} 1 & \text{when} \quad i = j \\ 0 & \text{when} \quad i \neq j \end{cases}. \tag{11.15}$$

The matrix elements of the identity matrix $\mathbf{I}$ are given by the Kronecker delta: $I_{ij} = \delta_{ij}$.

Problem b Show that expression (11.14) may be written in vector form as $\nabla p = \nabla\cdot(p\mathbf{I})$. Show that the law of conservation of momentum is given by

$$\frac{\partial(\rho\mathbf{v})}{\partial t} + \nabla\cdot\mathbf{S} = \mathbf{0}, \tag{11.16}$$

where $\mathbf{S}$ is defined by

$$\mathbf{S} = \rho\mathbf{v}\mathbf{v} + p\mathbf{I}. \tag{11.17}$$

This quantity describes the *radiation stress* [10] which accounts for the internal stress in an acoustic medium that is generated by the waves that propagate through the medium. Of course, when other external forces are present, they will lead to a right-hand side of (11.16) that is nonzero. This example shows that it can be arbitrary whether a certain physical effect is accounted for by a source term or by a current. This arbitrariness is caused by the fact that it is not clear what is internal to the system and what is external. In (11.13) the pressure force is treated as an external force whereas in (11.17) the pressure contributes to a current within the system. There is no objective reason to prefer one or the other of the two formulations.

You may find the inner products of vectors and the ∇-operator in expressions such as (11.12) confusing, and indeed a notation such as $\rho\mathbf{v}\cdot\nabla\mathbf{v}$ can be a source of error and confusion. When working with quantities like this it is clearer to explicitly write out the components of all vectors or tensors. In component form an equation such as (11.12) is written as:

$$\rho\frac{\partial v_i}{\partial t}+\sum_j \rho v_j \partial_j v_i = F_i . \tag{11.18}$$

Problem c Rewrite the continuity equation (11.6) in component form and redo the derivation of problem a with all equations in component form to arrive at the conservation law of momentum in component form:

$$\frac{\partial(\rho v_i)}{\partial t}+\sum_j \partial_j(\rho v_j v_i) = F_i . \tag{11.19}$$

In order to derive the law of energy conservation we start by deriving the conservation law for the kinetic energy (density)

$$E_K=\frac{1}{2}\rho v^2=\sum_i \frac{1}{2}\rho v_i v_i . \tag{11.20}$$

Problem d Express the partial time-derivative $\partial(\rho v^2)/\partial t$ in the time-derivatives $\partial(\rho v_i)/\partial t$ and $\partial v_i/\partial t$, use (11.18) and (11.19) to eliminate these time-derivatives and write the final results as:

$$\sum_i \frac{\partial(\frac{1}{2}\rho v_i v_i)}{\partial t}=-\sum_{i,j}\partial_j\left(\frac{1}{2}\rho v_i v_i v_j\right)+\sum_j v_j F_j . \tag{11.21}$$

Problem e Use definition (11.20) to rewrite the expression above as the conservation law of kinetic energy:

$$\frac{\partial E_K}{\partial t} + \nabla \cdot (\mathbf{v} E_K) = (\mathbf{v} \cdot \mathbf{F}). \tag{11.22}$$

This equation states that the kinetic energy current is given by $\mathbf{J} = \mathbf{v} E_K$, and this term describes how kinetic energy is transported by the flow. The term $(\mathbf{v} \cdot \mathbf{F})$ on the right-hand side denotes the source of kinetic energy. It was shown in Section 5.4 that $(\mathbf{v} \cdot \mathbf{F})$ is the power delivered by the force $\mathbf{F}$. This means that (11.22) states that the power produced by the force $\mathbf{F}$ is the source of kinetic energy.

In order to invoke the potential energy as well we assume for the moment that the force $\mathbf{F}$ is the gravitational force. Suppose there is a gravitational potential $V(\mathbf{r})$, then the gravitational force is given by

$$\mathbf{F} = -\rho \nabla V, \tag{11.23}$$

and the potential energy E_P is given by

$$E_P = \rho V. \tag{11.24}$$

Problem f Take the (partial) time-derivative of (11.24), use the continuity equation (11.6) to eliminate $\partial \rho / \partial t$, use that the potential V does not depend explicitly on time, and employ (11.23) and (11.24) to derive the conservation law of potential energy:

$$\frac{\partial E_P}{\partial t} + \nabla \cdot (\mathbf{v} E_P) = -(\mathbf{v} \cdot \mathbf{F}). \tag{11.25}$$

Note that this conservation law is very similar to the conservation law (11.22) for kinetic energy. The meaning of the second term on the left-hand side will be clear to you by now; it denotes the divergence of the current $\mathbf{v} E_P$ of potential energy. Note that the right-hand side of (11.24) has the opposite sign to the right-hand side of (11.22). This reflects the fact that when the force $\mathbf{F}$ acts as a source of kinetic energy, it acts as a sink of potential energy; the opposite signs imply that kinetic and potential energy are converted into each other. However, the total energy $E = E_K + E_P$ should have no source or sink.

Problem g Show that the total energy is source-free:

$$\frac{\partial E}{\partial t} + \nabla \cdot (\mathbf{v} E) = 0. \tag{11.26}$$

11.4 Heat equation

In the previous section we saw that the momentum and energy current could be derived from Newton's law. Such a rigorous derivation is not always possible. In this section the transport of heat is treated, and we will see that the law for heat transport cannot be derived rigorously. Consider the general conservation equation (11.3), where T is the temperature. (Strictly speaking we should derive the heat equation using a conservation law for the thermal energy rather than the temperature. The thermal energy is given by $\rho C T$, with C the heat capacity and ρ the mass density. When the specific heat and the density are constant, the distinction between thermal energy and temperature implies multiplication by a constant, but for simplicity this multiplication is left out here.)

The source term in the conservation equation is simply the amount of heat (normalized by the heat capacity) supplied to the medium. An example of such a source is the decay of radioactive isotopes that forms a major source in the heat budget of the Earth. The transport of heat is affected by the heat current **J**. In the Earth, heat can be transported by two mechanisms: heat conduction and heat advection. The first process is similar to the process of diffusion; it accounts for the fact that heat flows from warm regions to colder regions. The second process accounts for the heat that is transported by the flow field **v** in the medium. Therefore, the current **J** can be written as a sum of two components:

$$\mathbf{J} = \mathbf{J}^{conduction} + \mathbf{J}^{advection}. \tag{11.27}$$

The heat advection is given by

$$\mathbf{J}^{advection} = \mathbf{v}T, \tag{11.28}$$

which reflects that heat is simply carried around by the flow. This viewpoint of the process of heat transport is in fact too simplistic in many situations. Fletcher [42] describes how the human body loses heat during outdoor activities through four processes: conduction, advection, evaporation, and radiation. He describes in detail the conditions under which each of these processes dominate, and how the associated heat loss can be reduced. In the physics of the atmosphere, energy transport by radiation and by evaporation (or condensation) also plays a crucial role.

For the moment we focus on heat conduction. In general, heat flows from warm regions to colder regions. The vector ∇T points from cold regions to warmer regions. It is therefore logical that the heat conduction points in the opposite direction from the temperature gradient:

$$\mathbf{J}^{conduction} = -\kappa \nabla T, \tag{11.29}$$

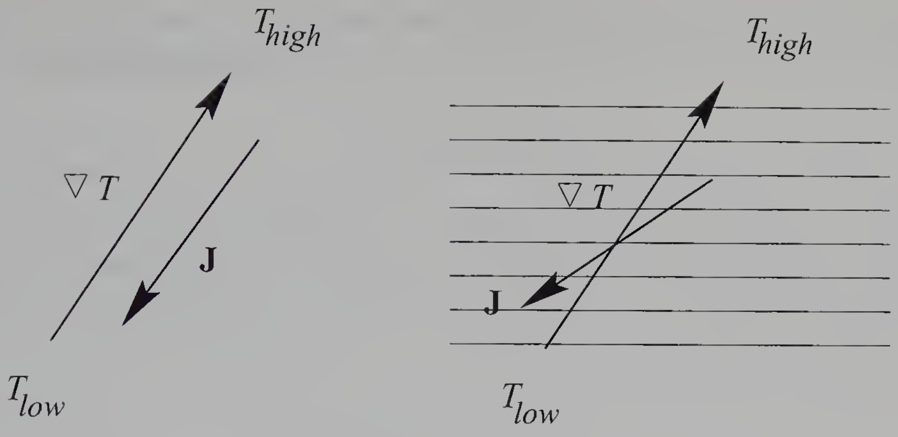

Fig. 11.1 Heat flow and temperature gradient in an isotropic medium (left-hand panel) and in a medium consisting of alternating layers of copper and styrofoam (right-hand panel).

see the left-hand panel of Figure 11.1. The constant κ is the heat conductivity. (For a given value of ∇T the heat conduction increases when κ increases, hence it indeed measures the conductivity.) However, the simple law (11.29) does not hold for every medium. Consider a medium consisting of alternating layers of a good heat conductor (such as copper) and a poor heat conductor (such as styrofoam). In such a medium the heat will be preferentially transported along the planes of the good heat conductor and the conductive heat flow $\mathbf{J}^{conduction}$ and the temperature gradient are not antiparallel, see the right-hand panel in Figure 11.1. In that case there is a matrix operator that relates $\mathbf{J}^{conduction}$ and ∇T: $J_i^{conduction} = -\sum_j \kappa_{ij}\partial_j T$, with κ_{ij} the heat conductivity tensor. In this section we restrict ourselves to the simple conduction law (11.29). Combining this law with the expressions (11.27), (11.28), and the conservation law (11.3) for heat gives:

$$\frac{\partial T}{\partial t} + \nabla \cdot (\mathbf{v}T - \kappa\nabla T) = S. \tag{11.30}$$

As a first example we consider a solid in which there is no flow ($\mathbf{v} = 0$). For a constant heat conductivity κ, (11.30) reduces to:

$$\frac{\partial T}{\partial t} = \kappa\nabla^2 T + S. \tag{11.31}$$

The expression is called the "heat equation", despite the fact that it holds only under special conditions. This expression is identical to Fick's law, which accounts for diffusion processes. This is not surprising since heat is transported by a diffusive process in the absence of advection.

We now consider heat transport in a one-dimensional medium (such as a bar) when there is no source of heat. In that case the heat equation reduces to

$$\frac{\partial T}{\partial t} = \kappa \frac{\partial^2 T}{\partial x^2}. \tag{11.32}$$

If we know the temperature throughout the medium at some initial time (i.e. $T(x, t = 0)$ is known), then (11.32) can be used to compute the temperature at later times. As a special case we consider a Gaussian-shaped temperature distribution at $t = 0$:

$$T(x, t = 0) = T_0 \exp\left(-\frac{x^2}{L^2}\right). \tag{11.33}$$

Problem a Sketch this temperature distribution and indicate the role of the constants T_0 and L.

We assume that the temperature profile maintains a Gaussian shape at later times but that the peak value and the width may change, that is we consider a solution of the following form:

$$T(x, t) = F(t)e^{-H(t)x^2}. \tag{11.34}$$

At this point the functions $F(t)$ and $H(t)$ are not yet known.

Problem b Show that these functions satisfy the initial conditions:

$$F(0) = T_0, \qquad H(0) = 1/L^2. \tag{11.35}$$

Problem c Show that for the special solution (11.34) the heat equation reduces to:

$$\frac{\partial F}{\partial t} - x^2 F \frac{\partial H}{\partial t} = \kappa \left(4FH^2x^2 - 2FH\right). \tag{11.36}$$

It is possible to derive equations for the time evolution of F and H by recognizing that (11.36) can only be satisfied for *all values of* x when all terms proportional to x^2 balance and when the terms independent of x balance.

Problem d Use this to show that $F(t)$ and $H(t)$ satisfy the following differential equations:

$$\frac{\partial F}{\partial t} = -2\kappa FH, \tag{11.37}$$

$$\frac{\partial H}{\partial t} = -4\kappa H^2. \tag{11.38}$$

It is easiest to solve the last equation first because it contains only $H(t)$ whereas (11.37) contains both $F(t)$ and $H(t)$.

Problem e Solve (11.38) with the initial condition (11.35) and show that:

$$H(t) = \frac{1}{4\kappa t + L^2}. \tag{11.39}$$

Problem f Solve (11.37) with the initial condition (11.35) and show that:

$$F(t) = T_0 \frac{L}{\sqrt{4\kappa t + L^2}}. \tag{11.40}$$

Inserting these solutions into (11.34) gives the temperature field at all times $t \geq 0$:

$$T(x,t) = T_0 \frac{L}{\sqrt{4\kappa t + L^2}} \exp\left(-\frac{x^2}{4\kappa t + L^2}\right). \tag{11.41}$$

Problem g Sketch the temperature for several later times and describe, using the solution (11.41), how the temperature profile changes as time progresses.

The total heat $Q^{total}(t)$ at time t is given by $Q^{total}(t) = \rho C \int_{-\infty}^{\infty} T(x,t)dx$, where C is the heat capacity, and ρ the mass density.

Problem h Show that the total heat does not change with time for the solution (11.41). Hint: reduce any integral of the form $\int_{-\infty}^{\infty} e^{-\alpha x^2} dx$ to the integral $\int_{-\infty}^{\infty} e^{-u^2} du$ with a suitable change of variables. You do not even have to use that $\int_{-\infty}^{\infty} e^{-u^2} du = \sqrt{\pi}$.

Problem i Show that for *any* solution of the heat equation (11.32), where the heat flux vanishes at the endpoints ($\kappa \partial_x T(x = \pm\infty, t) = 0$), the total heat $Q^{total}(t)$ is constant in time.

Problem j What happens to the special solution (11.41) when the temperature field evolves backward in time? Consider in particular times earlier than $t = -L^2/4\kappa$.

Problem k The peak value of the temperature field (11.41) decays as $1/\sqrt{4\kappa t + L^2}$ with time. Do you expect that in more dimensions this decay will be more rapid or slower with time? Do not do any calculations but use only your common sense!

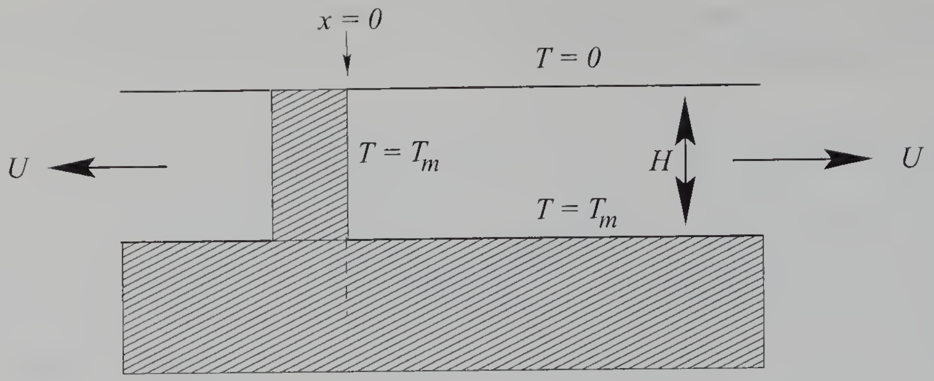

Fig. 11.2 Sketch of the cooling model of the oceanic lithosphere.

Up to this point, we have considered the conduction of heat in a medium without flow ($\mathbf{v} = 0$). In many applications the flow in the medium plays a crucial role in redistributing heat. This is particularly the case when heat is the source of convective motion, as for example in the Earth's mantle, the atmosphere, and the central heating systems in buildings. As an example of the role of advection we consider the cooling model of the oceanic lithosphere proposed by Parsons and Sclater [81].

At the mid-oceanic ridges, lithospheric material with thickness H is produced. At a ridge the temperature of this material is essentially the temperature T_m of mantle material. As shown in Figure 11.2, this implies that at $x = 0$ and at depth $z = H$ the temperature is given by the mantle temperature: $T(x = 0, z) = T(x, z = H) = T_m$. We assume that the velocity with which the plate moves away from the ridge is constant:

$$\mathbf{v} = U\hat{\mathbf{x}}. \tag{11.42}$$

We consider the situation in which the temperature is stationary. This does not imply that the flow vanishes; it means that the partial time-derivatives vanish: $\partial T/\partial t = 0$, $\partial \mathbf{v}/\partial t = 0$.

Problem 1 Show that in the absence of heat sources ($S = 0$) the conservation equation (11.30) reduces to:

$$U\frac{\partial T}{\partial x} = \kappa \left(\frac{\partial^2 T}{\partial x^2} + \frac{\partial^2 T}{\partial z^2} \right). \tag{11.43}$$

In general the thickness of the oceanic lithosphere is less than 100 km, whereas the width of ocean basins is several thousand kilometers.

Problem m Use this fact to explain that the following expression is a reasonable approximation to (11.43):

$$U\frac{\partial T}{\partial x}=\kappa\frac{\partial^2 T}{\partial z^2}. \tag{11.44}$$

Problem n Show that with the replacement $\tau = x/U$ this expression is identical to the heat equation (11.32).

Note that τ is the time it has taken the oceanic plate to move from its point of creation ($x = 0$) to the point under consideration (x), hence the time τ is simply the *age* of the oceanic lithosphere. This implies that solutions of the one-dimensional heat equation can be used to describe the cooling of oceanic lithosphere with the age of the lithosphere taken as the time variable. Accounting for cooling with such a model leads to a prediction of the depth of the ocean that increases as $\sqrt{t}$ with the age of the lithosphere. For ages less than about 100 million years this is in very good agreement with the observed ocean depth [81].

11.5 Explosion of a nuclear bomb

As an example of the use of conservation equations we study the condition under which a ball of uranium or plutonium can explode through a nuclear chain reaction. The starting point is once again the general conservation law (11.3), where Q is the concentration $N(\mathbf{r}, t)$ of neutrons per unit volume. We assume that the material is solid and that there is no flow: $\mathbf{v} = 0$. The neutron concentration is affected by two processes. First, the neutrons experience normal diffusion. For simplicity we assume that the neutron current is given by (11.29): $\mathbf{J} = -\kappa\nabla N$, with κ a constant. Second, neutrons are produced in the nuclear chain reaction. For example, when an atom of U_{235} absorbs one neutron, it may undergo fission and emit three free neutrons. This effectively constitutes a source of neutrons. The intensity of this source depends on the neutrons that are around to produce the fission of atoms. This implies that the source term is proportional to the neutron concentration: $S = \lambda N$, where λ is a positive constant that depends on the details of the nuclear reactions.

Problem a Show that the neutron concentration satisfies:

$$\frac{\partial N}{\partial t}=\kappa\nabla^2 N+\lambda N. \tag{11.45}$$

This equation needs to be supplemented with boundary conditions. We assume that the radioactive material is in the shape of a sphere of radius R. At the edge of the sphere the neutron concentration vanishes while at the center of the sphere the neutron concentration must remain finite for finite times:

$$N(r = R, t) = 0 \qquad \text{and} \qquad N(r = 0, t) \text{ is finite.} \tag{11.46}$$

We restrict our attention to solutions that are spherically symmetric: $N = N(r, t)$.

Problem b Apply separation of variables by writing the neutron concentration as $N(r, t) = F(r)H(t)$ and show that $F(r)$ and $H(t)$ satisfy the following equations:

$$\frac{\partial H(t)}{\partial t} = \mu H(t), \tag{11.47}$$

$$\nabla^2 F(r) + \frac{(\lambda - \mu)}{\kappa} F(r) = 0, \tag{11.48}$$

where μ is a separation constant that is not yet known.

Problem c Show that for positive μ there is an exponential growth of the neutron concentration with characteristic growth time $\tau = 1/\mu$.

Problem d Use the Laplacian in spherical coordinates to rewrite (11.48). Make the substitution $F(r) = f(r)/r$ and show that $f(r)$ satisfies:

$$\frac{\partial^2 f}{\partial r^2} + \frac{(\lambda - \mu)}{\kappa} f = 0. \tag{11.49}$$

Problem e Derive the boundary conditions at $r = 0$ and $r = R$ for $f(r)$.

Problem f Show that (11.49) with the boundary conditions derived in problem e can be satisfied only when

$$\mu = \lambda - \left(\frac{n\pi}{R}\right)^2 \kappa \qquad \text{for integer } n. \tag{11.50}$$

Problem g Show that for $n = 0$ the neutron concentration vanishes so that we only need to consider values $n \geq 1$.

Equation (11.50) gives the growth rate of the neutron concentration. It can be seen that the effects of unstable nuclear reactions and of neutron diffusion oppose each other. The λ term accounts for the growth of the neutron concentration through fission reactions: this term makes the inverse growth rate μ more positive. Conversely, the κ term accounts for diffusion: this term gives a negative contribution to μ.

Problem h What value of n gives the largest growth rate? Show that exponential growth of the neutron concentration (i.e. a nuclear explosion) can only occur when

$$R > \pi\sqrt{\frac{\kappa}{\lambda}}. \tag{11.51}$$

This implies that a nuclear explosion can only occur when the ball of fissionable material is larger than a certain critical size. If the ball is smaller than the critical size, more neutrons diffuse out of it than are created by fission, hence the nuclear reaction stops. In some of the earliest nuclear devices an explosion was created by bringing two half spheres that each were stable together to form one whole sphere that was unstable.

Problem i Suppose you have a ball of fissionable material that is just unstable and that you shape this material into a cube rather than a ball. Do you expect this cube to be stable or unstable? Do not use any equations!

11.6 Viscosity and the Navier–Stokes equation

Many fluids exhibit a certain degree of viscosity. In this section it will be shown that viscosity can be seen as an ad-hoc description of the momentum current in a fluid due to small-scale movements in the fluid. The starting point of the analysis is the equation of momentum conservation in a fluid:

$$\frac{\partial(\rho\mathbf{v})}{\partial t} + \nabla\cdot(\rho\mathbf{v}\mathbf{v}) = \mathbf{F} \tag{11.13}$$

In a real fluid, motion takes place at a large range of length scales from microscopic eddies to organized motions with a size comparable to the size of the fluid body. Whenever we describe a fluid, it is impossible to account for the motion at the very small length scales. This is not only so in analytical descriptions, but it is in particular the case in numerical simulations of fluid flow. For example, in current weather prediction schemes the motion of the air is computed on a grid with a distance of about 100 km between the grid points. When you look at the weather it is obvious that there is considerable motion at smaller length scales (e.g. cumulus clouds indicating convection, fronts, etc.). In general one cannot simply ignore the motion at these short length scales because these small-scale fluid motions transport significant amounts of momentum, heat, and other quantities such as moisture [78].

One way to account for the effect of the small-scale motion is to express the small-scale motion in the large-scale motion. It is not obvious that this is consistent with reality, but it appears to be the only way to avoid a complete description of the small-scale motion of the fluid (which would be impossible).

In order to do this, we assume that there is some length scale which separates the small-scale flow from the large-scale flow, and we decompose the velocity into a long-wavelength component $\mathbf{v}^L$ and a short-wavelength component $\mathbf{v}^S$:

$$\mathbf{v} = \mathbf{v}^L + \mathbf{v}^S. \tag{11.52}$$

In addition, we take spatial averages over a length scale that corresponds to the length scale that distinguishes the large-scale flow from the small-scale flow. This average is indicated by angle brackets: $\langle \cdots \rangle$. The average of the small-scale flow is zero, $\langle \mathbf{v}^S \rangle = 0$, while the average of the large-scale flow is equal to the large-scale flow, $\langle \mathbf{v}^L \rangle = \mathbf{v}^L$, because the large-scale flow by definition does not vary over the averaging length. For simplicity we assume that the density does not vary.

Problem a Use the expressions (11.13) and (11.52) to show that the momentum equation for the large-scale flow is given by:

$$\frac{\partial(\rho \mathbf{v}^L)}{\partial t} + \nabla \cdot (\rho \mathbf{v}^L \mathbf{v}^L) + \nabla \cdot (\langle \rho \mathbf{v}^S \mathbf{v}^S \rangle) = \mathbf{F}. \tag{11.53}$$

Show in particular why this expression contains a contribution that is quadratic in the small-scale flow, but that the terms that are linear in $\mathbf{v}^S$ do not contribute.

All the terms in (11.53) are familiar, except the last term on the left-hand side. This term exemplifies the effect of the small-scale flow on the large-scale flow since it accounts for the transport of momentum by the small-scale flow. It seems that at this point further progress is impossible without knowing the small-scale flow $\mathbf{v}^S$. One way to make further progress is to express the small-scale momentum current $\langle \rho \mathbf{v}^S \mathbf{v}^S \rangle$ in the large-scale flow.

Consider the large-scale flow shown in Figure 11.3. Whatever the small-scale motions are, in general they will have the character of mixing. In the example in the figure, the momentum is large at the top of the figure and smaller at the bottom. As a first approximation one may assume that the small-scale motions transport momentum in the direction opposite to the momentum gradient of the large-scale flow. By analogy with (11.29) we can approximate the momentum transport by the

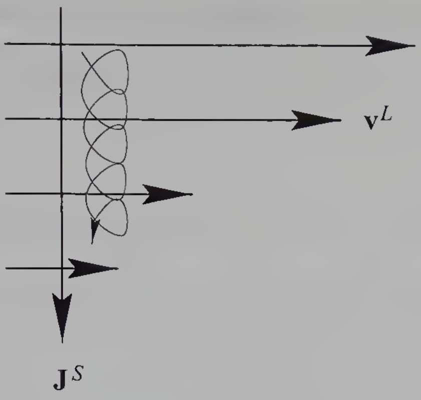

Fig. 11.3 The direction of momentum transport within a large-scale flow by small-scale motions.

small-scale flow by:

$$\mathbf{J}^S \equiv \left\langle \rho \mathbf{v}^S \mathbf{v}^S \right\rangle \approx -\mu \nabla \mathbf{v}^L, \tag{11.54}$$

where μ plays the role of a diffusion constant.

Problem b Insert this relation into (11.53), and drop the superscript L of $\mathbf{v}^L$ to show that large-scale flow satisfies:

$$\frac{\partial(\rho \mathbf{v})}{\partial t} + \nabla \cdot (\rho \mathbf{v}\mathbf{v}) = \mu \nabla^2 \mathbf{v} + \mathbf{F}. \tag{11.55}$$

This equation is called the Navier–Stokes equation. The first term on the right-hand side accounts for the momentum transport by small-scale motions. Effectively this leads to the viscosity of the fluid.

Problem c Viscosity tends to damp motion at smaller length scales more than motion at larger length scales. Show that the term $\mu \nabla^2 \mathbf{v}$ indeed affects shorter length scales more than larger length scales.

Problem d Do you think this treatment of the momentum flux due to small-scale motion is realistic? Can you think of an alternative?

Despite reservations that you may (or may not) have against the treatment of viscosity in this section, you should realize that the Navier–Stokes equation (11.55) is widely used in fluid mechanics.

11.7 Quantum mechanics and hydrodynamics

As we saw in Section 8.4, the behavior of atomic particles is described by Schrödinger's equation:

$$i\hbar\frac{\partial\psi(\mathbf{r},t)}{\partial t}=-\frac{\hbar^2}{2m}\nabla^2\psi(\mathbf{r},t)+V(\mathbf{r})\psi(\mathbf{r},t), \tag{8.14}$$

rather than Newton's law. In this section we reformulate the linear wave equation (8.14) as the laws of conservation of mass and momentum for a normal fluid. In order to do this we write the wavefunction ψ as

$$\psi=\sqrt{\rho}\times e^{(i/\hbar)\varphi}. \tag{11.56}$$

This equation is simply the decomposition of a complex function into its absolute value and its phase, hence ρ and φ are real functions. The factor $\hbar$ is a constant added here for notational convenience. This constant is characteristic of quantum mechanics and is called Planck's constant.

Problem a Insert the decomposition (11.56) into Schrödinger's equation (8.14), divide by $\sqrt{\rho}e^{(i/\hbar)\varphi}$ and separate the result into real and imaginary parts to show that ρ and φ satisfy the following differential equations:

$$\frac{\partial\rho}{\partial t}+\nabla\cdot\left(\rho\frac{1}{m}\nabla\varphi\right)=0, \tag{11.57}$$

$$\frac{\partial\varphi}{\partial t}+\frac{1}{2m}\left|\nabla\varphi\right|^2+\frac{\hbar^2}{8m}\left(\frac{1}{\rho^2}\left|\nabla\rho\right|^2-\frac{2}{\rho}\nabla^2\rho\right)=-V. \tag{11.58}$$

The problem is that at this point we do not yet have a velocity. Let us define the following velocity vector:

$$\mathbf{v}\equiv\frac{1}{m}\nabla\varphi. \tag{11.59}$$

Problem b Show that this definition of the velocity is identical to the velocity obtained in (8.21) for a plane wave.

Problem c Show that with this definition of the velocity, (11.57) is identical to the continuity equation:

$$\frac{\partial\rho}{\partial t}+\nabla\cdot(\rho\mathbf{v})=0. \tag{11.6}$$

Problem d In order to reformulate (11.58) as an equation of conservation of momentum, differentiate (11.58) with respect to x_i. Then use the definition (11.59) and the relation between force and potential ($\mathbf{F} = -\nabla V$) to write the result as:

$$\frac{\partial v_i}{\partial t} + \frac{1}{2}\sum_j \partial_i (v_j v_j) + \frac{\hbar^2}{8m}\left[\partial_i \left(\frac{1}{\rho^2}|\nabla \rho|^2\right) - 2\partial_i \left(\frac{1}{\rho}\nabla^2 \rho\right)\right] = \frac{1}{m}F_i. \tag{11.60}$$

The second term on the left-hand side does not look very much like the term $\sum_j \partial_j(\rho v_j v_i)$ in the left-hand side of (11.13). To make progress we need to rewrite the term $\sum_j \partial_i(v_j v_j)$ as a term of the form $\sum_j \partial_j(v_j v_i)$. In general these terms are different.

Problem e Show that for the special case that the velocity is the gradient of a scalar function (as in expression (11.59)):

$$\sum_j \frac{1}{2}\partial_i(v_j v_j) = \sum_j \partial_j(v_j v_i). \tag{11.61}$$

With this step we can rewrite the second term on the left-hand side of (11.60). Part of the third term in (11.60) we will designate as Q_i:

$$Q_i \equiv -\frac{1}{8}\left[\partial_i \left(\frac{1}{\rho^2}|\nabla \rho|^2\right) - 2\partial_i \left(\frac{1}{\rho}\nabla^2 \rho\right)\right]. \tag{11.62}$$

Problem f Use (11.6) and (11.60)–(11.62) to derive that:

$$\frac{\partial(\rho \mathbf{v})}{\partial t} + \nabla \cdot (\rho \mathbf{v}\mathbf{v}) = \frac{\rho}{m}\left(\mathbf{F} + \hbar^2 \mathbf{Q}\right). \tag{11.63}$$

Note that this equation is identical to the momentum equation (11.13). This implies that the Schrödinger equation is equivalent to the continuity equation (11.6) and to the momentum equation (11.13) for a classical fluid. In Section 8.4 we saw that atomic particles behave as waves rather than as point-like particles. In this section we have discovered that these particles also behave like a fluid! This has led to hydrodynamic formulations of quantum mechanics [48, 66]. In general, quantum-mechanical phenomena depend critically on Planck's constant. Quantum mechanics reduces to classical mechanics in the limit $\hbar \to 0$. The only place where Planck's constant occurs in (11.63) is in the additional force: $\mathbf{Q}$ multiplied by the square of Planck's constant. This implies that the action of the force term **Q** is fundamentally quantum-mechanical; it has no analogue in classical mechanics.

Problem g Suppose we consider a particle in one dimension that is represented by the following wave function:

$$\psi(x,t) = e^{-x^2/L^2} e^{i(kx-\omega t)}. \tag{11.64}$$

Sketch the corresponding probability-density ρ and use (11.62) to deduce that the quantum force Q acts to broaden the wavefunction with time.

This example shows that (at least for this case) the quantum force **Q** makes the wavefunction "spread out" with time. This reflects the fact that if a particle propagates with time, its position becomes more and more uncertain.

The acoustic wave equation (8.7) cannot be transformed into the continuity equation (11.6) and the momentum equation (11.13), despite the fact that these equations describe the conservation of mass and momentum in an acoustic medium. The reason for this paradox is that in the derivation of the acoustic wave equation from the continuity equation and the momentum equation, the advective terms $\nabla \cdot (\rho \mathbf{v})$ and $\nabla \cdot (\rho \mathbf{v}\mathbf{v})$ have been ignored. Once these terms have been ignored there is no transformation of variables that can bring them back. This contrasts the acoustic wave equation from the Schrödinger equation, which implicitly retains the physics of the advection of mass and momentum.

12

Scale analysis

In most situations, the equations that we would like to solve in mathematical physics are too complicated to solve analytically. One of the reasons for this is often that an equation contains many different terms which make the problem simply too complex to be manageable. However, many of these terms may in practice be very small. Ignoring these small terms can simplify the problem to such an extent that it can be solved in closed form. Moreover, by deleting terms that are small one is able to focus on the terms that are significant and that contain the relevant physics. In this sense, ignoring small terms can actually give a better physical insight into the processes that really do matter.

Scale analysis is a technique in which one estimates the different terms in an equation by considering the scale over which the relevant parameters vary. This is an extremely powerful tool for simplifying problems. A comprehensive overview of this technique with many applications is given by Kline [56] and in Chapter 6 of Lin *et al.* [63]. Interesting examples of the application of scaling arguments to biology are given by Vogel [113].

With the application of scale analysis one caveat must be made. One of the major surprises of classical physics of the twentieth century was the discovery of chaos in dynamical systems [104]. In a chaotic system small changes in the initial conditions lead to a change in the time evolution of the system that grows exponentially with time. Deleting small terms from the equation of motion of such a system can have a similar effect; the effects of omitting small terms can lead to changes in the system that may grow exponentially with time. This means that for chaotic systems one must be careful in omitting small terms from the equations.

The principle of scale analysis is first applied to the problem of determining the sense of rotation of a vortex in an emptying bathtub. Many of the equations that are used in physics are differential equations. For this reason it is crucial in scale analysis to be able to estimate the order of magnitude of derivatives. The estimation of derivatives is therefore treated in Section 12.2. In subsequent sections this is then applied to a variety of different problems.

12.1 Vortex in a bathtub

Often when you empty a bathtub you can observe a beautiful vortex above the drain in the bathtub. When you ask people about the sense of rotation of this vortex many people will reply that the sense of rotation is in the counterclockwise direction on the northern hemisphere and in the clockwise direction in the southern hemisphere (or vice versa). This statement is wrong and the easiest way to verify this is to take a bath, empty the bathtub and observe that it is as easy to obtain a vortex that rotates in the clockwise direction as it is to create a vortex that rotates in the counterclockwise direction. In fact, if you have ever had the opportunity to take a bath in Nairobi, you would have had the chance to observe that on the equator (where the effective rotation of the Earth vanishes) a bathtub drains in exactly the same way as at higher latitudes.

Yet when you ask your friends you may find that it is very difficult to convince them of these facts. In that case you may be able to convince them with a calculation. One approach would be to take the equations of fluid flow on a rotating Earth and compute numerically in all details the flow in your bathtub. However, this approach is not only terribly impractical but it is also very difficult to carry out because you know neither the detailed shape of your bathtub nor the initial conditions of the water in the bathtub before you drain it. An even more serious drawback of this approach is that this numerical simulation does not give much physical insight. At best it gives a perfect simulation of the real bathtub, but in that case it would be better to do the experiment in the real bathtub.

As an alternative we can estimate the relative strength of the forces that are acting on the fluid in the vortex. In Figure 12.1 the forces are shown that act on a fluid parcel in a vortex on the Earth that rotates with angular velocity Ω. The dominant forces are the pressure force $\mathbf{F}_{pres}$ that is associated with the pressure gradient in the fluid, the Coriolis force $\mathbf{F}_{Cor}$ due to the rotation of the Earth, and the centrifugal force $\mathbf{F}_{cent}$ that is due to the circular motion of the fluid in the vortex. The rotational velocity of the vortex is denoted by ω. The only force that depends on the position on the Earth is the Coriolis force $\mathbf{F}_{Cor}$. In Figure 12.1 this force points to the right-hand side of the velocity vector; this is the case in the northern hemisphere whereas in the southern hemisphere this force would point to the left-hand side of the velocity vector. The centrifugal force always points outward from the vortex while the pressure force always points into the vortex. This means that any asymmetry between the behavior of a vortex in the northern hemisphere and one in the southern hemisphere must be due to the Coriolis force.

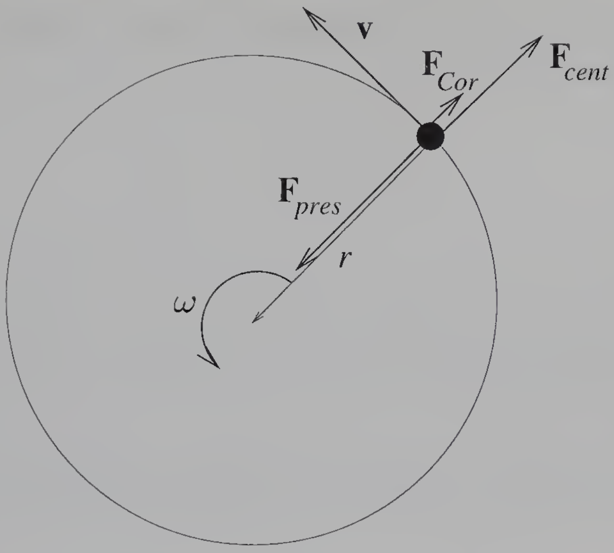

Fig. 12.1 The forces that act in a fluid vortex on the northern hemisphere of a rotating Earth.

In order to test our hypothesis we need to estimate the strength of the Coriolis force compared to the other forces that are operative. In the balance of forces, the pressure force is balanced by the Coriolis force, or by the centrifugal force or by both of them. For this reason, it is sufficient to estimate the strength of the Coriolis force compared to the centrifugal force. We derive in Section 13.3 that the Coriolis force is given by $\mathbf{F}_{Cor} = -2\Omega \times \mathbf{v}$, where Ω is the rotation vector of the Earth's rotation. Since we are only estimating orders of magnitude this means that the Coriolis force is of the order

$$F_{Cor} \sim \Omega v. \tag{12.1}$$

The strength of the centrifugal force is given by

$$F_{cent} = \frac{v^2}{r}. \tag{12.2}$$

Problem a Use Figure 12.1 to deduce that $v/r = \omega$ and then use this result to show that the ratio of the Coriolis force to the centrifugal force is approximately given by

$$\frac{F_{Cor}}{F_{cent}} \sim \frac{\Omega}{\omega}. \tag{12.3}$$

The ratio of the Coriolis force to the centrifugal force is thus of the order of the ratio of the rotation rate of the Earth to the rotation rate of the vortex in the bathtub.

Problem b Assuming that the vortex rotates once a second, use the previous expression to show that

$$\frac{F_{Cor}}{F_{cent}} \sim 10^{-5}. \tag{12.4}$$

This means that the Coriolis force is much smaller than all other forces that are operating on the fluid parcel. The asymmetry in the balance of forces created by the Coriolis force is in practice also much smaller than the asymmetry in the shape of the bathtub and the asymmetry in the initial conditions of the fluid motion before you drain the bathtub. We can conclude from this that the Earth's rotation is negligible in the dynamics of the vortex; hence the sense of rotation of a vortex in a bathtub does not depend on the geographic location of the bathtub. This example shows that one can sometimes learn more from simple physical arguments and estimates of orders of magnitude than from very complex numerical calculations.

It follows from (12.3) that the centrifugal force and the Coriolis force are of the same order of magnitude when the rotation rate of the fluid in the bathtub is of the same order of magnitude as the Earth's rotation.

Problem c Estimate the ratio of the Coriolis force to the centrifugal force for the atmospheric motion around a depression.

You will have found that for the atmosphere the Coriolis force *is* one of the dominant forces. This is the reason that in the northern hemisphere the air moves in the counterclockwise direction around a low-pressure area and in the clockwise direction around a high-pressure center (and vice versa on the southern hemisphere). A low-pressure area is indeed very similar to the vortex in a bathtub. This analogy is in fact the cause of the misconception that the vortex in a bathtub always turns in one direction depending on geographic location. The major difference is, however, that the relative strengths of the forces that are operative in the atmosphere and in a bathtub are completely different. This result has been derived here with a simple scale analysis.

12.2 Three ways to estimate a derivative

In this section three different ways to estimate the derivative of a function $f(x)$ are derived. The *first* way to estimate the derivative is to realize that the derivative is nothing but the slope of the function $f(x)$. Consider Figure 12.2 in which the function $f(x)$ is assumed to be known at neighboring points x and $x + h$.

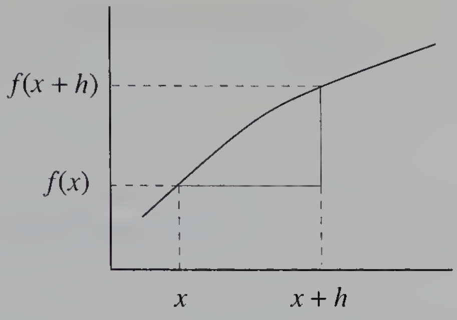

Fig. 12.2 The slope of a function $f(x)$ that is known at positions x and $x + h$.

Problem a Deduce from the geometry of this figure that the slope of the function at x is approximately given by $[f(x+h) - f(x)]/h$.

Since the slope is the derivative this means that the derivative of the function is approximately given by

$$\frac{df}{dx} \approx \frac{f(x+h) - f(x)}{h} . \tag{12.5}$$

The *second* way to derive the same result is to realize that the derivative is defined by the following limit:

$$\frac{df}{dx} \equiv \lim_{h\to 0} \frac{f(x+h) - f(x)}{h} . \tag{12.6}$$

If we consider the right-hand side of this expression without taking the limit, we do not quite obtain the derivative, but as long as h is sufficiently small we obtain the approximation (12.5).

The problem with estimating the derivative of $f(x)$ in the previous ways is that although we obtain an estimate of the derivative, we do not know how good these estimates are. We know that if $f(x)$ was a straight line, which has a constant slope, the estimate (12.5) would be exact. Hence it is the deviation of $f(x)$ from a straight line that makes (12.5) only an approximation. This means that it is the curvature of $f(x)$ that accounts for the error in the approximation (12.5). The *third* way of estimating the derivative provides this error estimate as well.

Problem b Consider the Taylor series (3.17). Truncate this series after the second order term and solve the resulting expression for df/dx to derive that

$$\frac{df}{dx} = \frac{f(x+h) - f(x)}{h} - \frac{1}{2}\frac{d^2 f}{dx^2} h + \cdots , \tag{12.7}$$

where the dots indicate terms of order h^2 and higher.

In the limit $h \to 0$ the last term vanishes and (12.6) is obtained. When one ignores the last term in (12.7) for finite h one once again obtains the approximation (12.5).

Problem c Use (12.7) to show that the error made in the approximation (12.5) indeed depends on the *curvature* of the function $f(x)$.

The approximation (12.5) has a variety of applications. The first is the numerical solution of differential equations. Suppose one has a differential equation that cannot be solved in closed form. To fix our minds, consider the differential equation

$$\frac{df}{dx} = G(f(x), x), \tag{12.8}$$

with initial value

$$f(0) = f_0. \tag{12.9}$$

If this equation cannot be solved in closed form, one can solve it numerically by evaluating the function $f(x)$ not for every value of x, but only at a finite number of x-values that are separated by a distance h. These points x_n are given by $x_n = nh$, and the function $f(x)$ at location x_n is denoted by f_n:

$$f_n \equiv f(x_n). \tag{12.10}$$

Problem d Show that the derivative df/dx at location x_n can be approximated by:

$$\frac{df}{dx}(x_n) = \frac{1}{h}(f_{n+1} - f_n). \tag{12.11}$$

Problem e Insert this result into the differential equation (12.8) and solve the resulting expression for f_{n+1} to show that:

$$f_{n+1} = f_n + hG(f_n, x_n). \tag{12.12}$$

This is all we need to numerically solve the differential equation (12.8) with the boundary condition (12.9). Once f_n is known, (12.12) can be used to compute f_{n+1}. This means that the function can be computed at all values of the grid points x_n recursively. To start this process, one uses the boundary condition (12.9) that gives the value of the function at location $x_0 = 0$. This technique for estimating the derivative of a function can be extended to higher order derivatives as well so that second order differential equations can also be solved numerically. In practice, one has to pay serious attention to the *stability* of the numerical solution.

The requirements of stability and numerical efficiency have led to many refinements of the numerical methods for solving differential equations. The interested reader can consult Press *et al.* [84] for an introduction and many practical algorithms.

The estimate (12.5) has a second important application because it allows us to estimate the order of magnitude of a derivative. Suppose a function $f(x)$ varies over a characteristic range of values F and that this variation takes place over a characteristic distance L. It follows from (12.5) that the derivative of $f(x)$ is of the order of the ratio of the variation of the function $f(x)$ divided by the length scale over which the function varies. In other words:

$$\left|\frac{df}{dx}\right| \approx \frac{\text{variation of the function } f(x)}{\text{length scale of the variation}} \sim \frac{F}{L}. \tag{12.13}$$

In this expression the term $\sim F/L$ indicates that the derivative is of the order F/L. Note that this is not in general an accurate estimate of the precise value of the function $f(x)$, it only provides us with an estimate of the order of magnitude of a derivative. However, this is all we need to carry out scale analysis.

Problem f Suppose $f(x)$ is a sinusoidal wave with amplitude A and wavelength λ:

$$f(x) = A\sin\left(\frac{2\pi x}{\lambda}\right). \tag{12.14}$$

Show that (12.13) implies that the order of magnitude of the derivative of this function is given by $|df/dx| \sim O(A/\lambda)$. Compare this estimate of the order of magnitude with the true value of the derivative and pay attention both to the numerical value as well as to the spatial variation.

From the previous estimate we can learn two things. First, the estimate (12.13) is only a rough estimate that *locally* can be very poor. One should always be aware that (12.13) may break down at certain points and that this can cause errors in the subsequent scale analysis. Second, (12.13) differs by a factor 2π from the true derivative. However, $2\pi \approx 6.28$, which is not a small number (compared to 1). Therefore you must be aware that hidden numerical factors may enter scaling arguments.

12.3 Advective terms in the equation of motion

As a first example of scale analysis we consider the role of advective terms in the equation of motion. As shown in (11.12) the equation of motion for a continuous medium is given by

$$\frac{\partial \mathbf{v}}{\partial t} + \mathbf{v} \cdot \nabla \mathbf{v} = \frac{1}{\rho} \mathbf{F}. \tag{12.15}$$

Note that we have divided by the density compared to the original expression (11.12). This equation can describe the propagation of acoustic waves when **F** is the pressure force, and it accounts for elastic waves when **F** is given by the elastic forces in the medium. We consider the situation in which waves with a wavelength λ and a period T propagate through the medium.

The advective terms $\mathbf{v} \cdot \nabla \mathbf{v}$ often pose a problem in solving this equation. This is because the partial time-derivative $\partial \mathbf{v}/\partial t$ is *linear* in the velocity **v** but the advective terms $\mathbf{v} \cdot \nabla \mathbf{v}$ are *nonlinear* in the velocity **v**. Since linear equations are in general much easier to solve than nonlinear equations it is useful to know under which conditions the advective terms $\mathbf{v} \cdot \nabla \mathbf{v}$ can be ignored compared with the partial derivative $\partial \mathbf{v}/\partial t$.

Problem a Let the velocity of the continuous medium have a characteristic value V. Show that $|\partial \mathbf{v}/\partial t| \sim V/T$ and that $|\mathbf{v} \cdot \nabla \mathbf{v}| \sim V^2/\lambda$.

Problem b Show that this means that the ratio of the advective terms to the partial time-derivative is given by

$$\frac{|\mathbf{v} \cdot \nabla \mathbf{v}|}{|\partial \mathbf{v}/\partial t|} \sim \frac{V}{c}, \tag{12.16}$$

where $c = \lambda/T$ is the phase velocity with which the waves propagate through the medium.

This result implies that the advective terms can be ignored when the velocity of the medium itself is much less than the velocity of the waves propagating through the medium:

$$V \ll c. \tag{12.17}$$

In other words, when the amplitude of the wave motion is so small that the velocity of the particle motion in the medium is much less than the phase velocity of the waves, one can ignore the advective terms in the equation of motion.

Problem c Suppose an earthquake causes a ground displacement of 1 mm at a frequency of 1 Hz at a large distance. The wave velocity of seismic P-waves is of the order of 5 km/s near the surface. Show that in that case $V/c \sim 10^{-6}$.

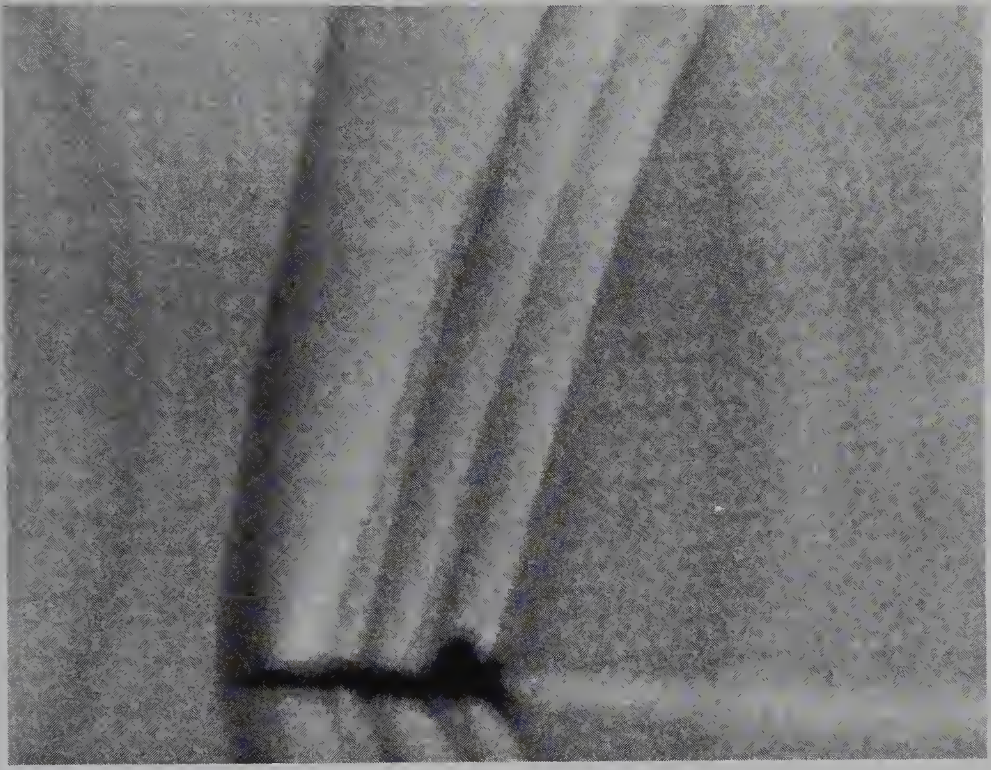

Fig. 12.3 The shock waves generated by a T38 flying at Mach 1.1 (a speed of 1.1 times the speed of sound) as made visible with the schlieren method.

The small value of V/c implies that for the propagation of elastic waves due to earthquakes one can ignore advective terms in the equation of motion. Note, however, that this is not necessarily true near the earthquake where the motion is much more violent and where the associated velocity of the rocks is not necessarily much smaller than the wave velocity.

There are a number of physical phenomena that are intimately related to the presence of the advective terms in the equation of motion. One important phenomenon is the occurrence of shock waves when the motion of the medium exceeds the wave velocity. A prime example of shock waves is the sonic boom made by an aircraft that moves at a velocity greater than the speed of sound [55, 107]. A spectacular example can be seen in Figure 12.3 where the shock waves generated by a T38 flying at a speed of Mach 1.1 at an altitude of 13 700 ft can be seen. These shock waves are visualized using the schlieren method [61] which is an optical technique to convert phase differences of light waves into amplitude differences.

Another example of shock waves is the formation of the *hydraulic jump*. You may not know what a hydraulic jump is, but you have surely seen one! Consider water flowing down a channel such as a mountain stream as shown in Figure 12.4. The flow velocity is denoted by v. At the bottom of the channel a rock disrupts the flow. This rock generates water waves that propagate with a velocity c compared to

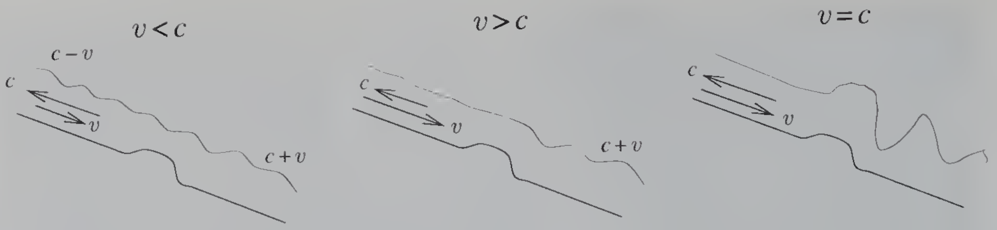

Fig. 12.4 Waves on a water flowing over a rock when $v < c$ (left-hand panel), $v > c$ (middle panel), and $v = c$ (right-hand panel).

the moving water. When the flow velocity is less than the wave velocity ($v < c$, see the left-hand panel of Figure 12.4) the waves propagate upstream with an absolute velocity $c - v$ and propagate downstream with an absolute velocity $c + v$. When the flow velocity is larger than the wave velocity ($v > c$, see the middle panel of Figure 12.4) the waves move downstream only because the wave velocity is not sufficiently large to move the waves against the current. The most interesting case is when the flow velocity equals the wave velocity ($v = c$, see the right-hand panel of Figure 12.4). In that case the waves that move upstream have an absolute velocity given by $c - v = 0$. In other words, these waves do not move with respect to the rock that generates the waves. This wave is continuously excited by the rock, and through a process similar to an oscillator that is driven at its resonance frequency the wave grows and grows until it ultimately breaks and becomes turbulent. This is the reason why one sees strong turbulent waves over boulders and other irregularities in streams. For further details on channel flow and hydraulic jumps the reader can consult Chapter 9 of Whitaker [118]. In general, the advective terms play a crucial role in the steepening and breaking of waves and in the formation of shock waves. This is described in much detail by Whitham [119].

12.4 Geometric ray theory

Geometric ray theory is an approximation that accounts for the propagation of waves along lines through space. The theory finds its conceptual roots in optics, where for a long time one has observed that a light beam propagates along a well-defined trajectory through lenses and other optical devices. Mathematically, this behavior of waves is accounted for in geometric ray theory, or more briefly "ray theory."

Ray theory is derived here for the acoustic wave equation rather than for the propagation of light because pressure waves are described by a scalar equation rather than the vector equation that governs the propagation of electromagnetic waves. The starting point is the acoustic wave equation (8.7):

$$\rho\nabla\cdot\left(\frac{1}{\rho}\nabla p\right)+\frac{\omega^2}{c^2}p=0. \tag{12.18}$$

For simplicity the source term on the right-hand side has been set to zero. In addition, the relation $c^2 = \kappa/\rho$ has been used to eliminate the bulk modulus κ in favor of the wave velocity c. Both the density ρ and the wave velocity are arbitrary functions of space.

In general it is not possible to solve this differential equation in closed form. Instead we seek an approximation by writing the pressure as:

$$p(\mathbf{r}, \omega) = A(\mathbf{r}, \omega) e^{i\psi(\mathbf{r}, \omega)}, \tag{12.19}$$

with A and ψ real functions. Any function $p(\mathbf{r}, \omega)$ can be written in this way.

Problem a Insert the solution (12.19) in the acoustic wave equation (12.18), separate the real and imaginary parts of the resulting equation to deduce that (12.18) is equivalent to the following equations:

$$\underbrace{\nabla^2 A}_{(1)} - \underbrace{A\,|\nabla\psi|^2}_{(2)} - \underbrace{\frac{1}{\rho}(\nabla\rho \cdot \nabla A)}_{(3)} + \underbrace{\frac{\omega^2}{c^2} A}_{(4)} = 0, \tag{12.20}$$

and

$$2\,(\nabla A \cdot \nabla\psi) + A\nabla^2\psi - \frac{1}{\rho}(\nabla\rho \cdot \nabla\psi)\,A = 0. \tag{12.21}$$

These equations are even harder to solve than the acoustic wave equation because they are nonlinear in the unknown functions A and ψ whereas the acoustic wave equation is linear in the pressure p. However, (12.20) and (12.21) form a good starting point for making the ray-geometric approximation. First we analyze (12.20).

Assume that the density varies on a length scale L_ρ, and that the amplitude A of the wave field varies on a characteristic length scale L_A. Furthermore the wavelength of the waves is denoted by λ.

Problem b Explain that the wavelength is the length scale over which the phase ψ of the waves varies.

Problem c Use the results of Section 12.2 to obtain the following estimates of the order of magnitude of the terms (1)–(4) in equation (12.20):

$$\left.\begin{array}{ll} \left|\nabla^2 A\right| \sim \dfrac{A}{L_A^2}, & A\,|\nabla\psi|^2 \sim \dfrac{A}{\lambda^2}, \\ \left|\dfrac{1}{\rho}(\nabla\rho \cdot \nabla A)\right| \sim \dfrac{A}{L_A L_\rho}, & \dfrac{\omega^2}{c^2} A \sim \dfrac{A}{\lambda^2}. \end{array}\right\} \tag{12.22}$$

To make further progress we assume that the length scales of both the density variations and the amplitude variations are much longer than a wavelength: $\lambda \ll L_A$ and $\lambda \ll L_\rho$.

Problem d Show that under this assumption terms (1) and (3) in equation (12.20) are much smaller than terms (2) and (4).

Problem e Convince yourself that ignoring terms (1) and (3) in (12.20) gives the following (approximate) expression:

$$|\nabla\psi|^2 = \frac{\omega^2}{c^2}. \tag{12.23}$$

Problem f The approximation (12.23) was obtained under the premise that $|\nabla\psi| \sim 1/\lambda$. Show that this assumption is satisfied by the function ψ in (12.23).

Whenever one makes approximations by deleting terms that scale analysis predicts to be small, one has to check that the final solution is consistent with the scale analysis that is used to derive the approximation.

Note that the original equation (12.20) contains both the amplitude A and the phase ψ but that (12.23) contains the phase only. The approximation that we have made has thus decoupled the phase from the amplitude; this simplifies the problem considerably. The frequency enters the right-hand side of this equation only through a simple multiplication with ω^2. The frequency dependence of ψ can be found by substituting

$$\psi(\mathbf{r}, \omega) = \omega\tau(\mathbf{r}). \tag{12.24}$$

Problem g Show that after this substitution (12.23) and (12.21) are given by

$$|\nabla\tau(\mathbf{r})|^2 = \frac{1}{c^2}, \tag{12.25}$$

and

$$2(\nabla A \cdot \nabla\tau) + A\nabla^2\tau - \frac{1}{\rho}(\nabla\rho \cdot \nabla\tau)A = 0. \tag{12.26}$$

According to (12.25) the function $\tau(\mathbf{r})$ does not depend on frequency. Note that (12.26) for the amplitude does not contain any frequency dependence either. This means that the amplitude also does not depend on frequency: $A = A(\mathbf{r})$. This has important consequences for the shape of the wave field in the ray-geometric approximation. Suppose that the wave field is excited by a source function $s(t)$ in the time domain that is represented in the frequency domain by a complex function $S(\omega)$. (The forward and backward Fourier transform is defined by (15.42) and (15.43).) In the frequency domain the response is given by (12.19) multiplied with the source function $S(\omega)$. Using that A and τ do not depend on frequency, the pressure in the time domain can be written as:

$$p(\mathbf{r}, t) = \int_{-\infty}^{\infty} A(\mathbf{r}) e^{i\omega\tau(\mathbf{r})} e^{-i\omega t} S(\omega)\, d\omega. \tag{12.27}$$

Problem h Use this expression to show that the pressure in the time domain can be written as:

$$p(\mathbf{r}, t) = A(\mathbf{r}) s(t - \tau(\mathbf{r})). \tag{12.28}$$

This is an important result because it implies that the time dependence of the wave field is everywhere given by the same source-time function $s(t)$. In a ray-geometric approximation the shape of the waveforms is everywhere the same. There are no frequency-dependent effects in a ray-geometric approximation.

Problem i Explain why this implies that geometric ray theory cannot be used to explain why the sky is blue.

The absence of any frequency-dependent wave propagation effects is both the strength and the weakness of ray theory. It is a strength because the wave fields can be computed in a simple way once $\tau(\mathbf{r})$ and $A(\mathbf{r})$ are known. The theory also tells us that this is an adequate description of the wave field as long as the frequency is sufficiently high that $\lambda \ll L_A$ and $\lambda \ll L_\rho$. However, many wave propagation phenomena are in practice frequency-dependent, and it is the weakness of ray theory that it cannot account for these phenomena.

According to (12.28) the function $\tau(\mathbf{r})$ accounts for the time delay of the waves to travel to the point $\mathbf{r}$. Therefore, $\tau(\mathbf{r})$ is the *travel time* of the wave field. The travel time is described by the differential equation (12.25); this equation is called the *eikonal equation*.

Problem j Show that it follows from the eikonal equation that $\nabla\tau$ can be written as:

$$\nabla\tau = \hat{\mathbf{n}}/c, \tag{12.29}$$

where $\hat{\mathbf{n}}$ is a unit vector. Show also that $\hat{\mathbf{n}}$ is perpendicular to the surface $\tau = const.$

The vector $\hat{\mathbf{n}}$ defines the direction of the *rays* along which the wave energy propagates through the medium. Taking suitable derivatives of (12.29) one can derive the *equation of kinematic ray-tracing*. This is a second order differential equation for the position of the rays; details are given by Virieux [112] or Aki and Richards [3].

Once $\tau(\mathbf{r})$ is known, one can compute the amplitude $A(\mathbf{r})$ from (12.26). We have not yet applied any scale analysis to this expression. We will not do so, because it can be solved exactly. Let us first simplify this differential equation by considering the dependence on the density ρ in more detail.

Problem k Write $A = \rho^{\alpha} B$, where the constant α is not yet determined. Show that the transport equation results in the following differential equation for $B(\mathbf{r})$:

$$(2\alpha - 1)(\nabla\rho \cdot \nabla\tau) B + 2\rho(\nabla B \cdot \nabla\tau) + \rho B \nabla^2 \tau = 0. \tag{12.30}$$

Choose the constant α in such a way that the gradient of the density disappears from the equation and show that the remaining terms can be written as $\nabla \cdot \left(B^2 \nabla\tau\right) = 0$. Finally show using (12.29) that this implies the following differential equation for the amplitude:

$$\nabla \cdot \left(\frac{1}{\rho c} A^2 \hat{\mathbf{n}}\right) = 0. \tag{12.31}$$

Equation (12.31) states that the divergence of the vector $\left(A^2/\rho c\right)\hat{\mathbf{n}}$ vanishes, hence the flux of this vector through any closed surface that does not contain the source of the wave-field vanishes, see Section 8.1. This is not surprising, because the vector $\left(A^2/\rho c\right)\hat{\mathbf{n}}$ accounts for the energy flux of acoustic waves. Expression (12.31) implies that the net flux of this vector through any closed surface is equal to zero. This means that all the energy that flows into the surface must also flow out through the surface again. The transport equation in the form (12.31) is therefore a statement of energy conservation. Virieux [112] or Aki and Richards [3] show how one can compute this amplitude once the location of the rays is known.

An interesting complication arises when the energy is focused at a point or on a surface in space. Such an area of focusing is called a *caustic*. A familiar example of a caustic is the rainbow. One can show that at a caustic, the ray-geometric approximation leads to an infinite amplitude of the wave field [112].

Problem l Show that when the amplitude becomes infinite in a finite region of space the condition $\lambda \ll L_A$ must be violated.

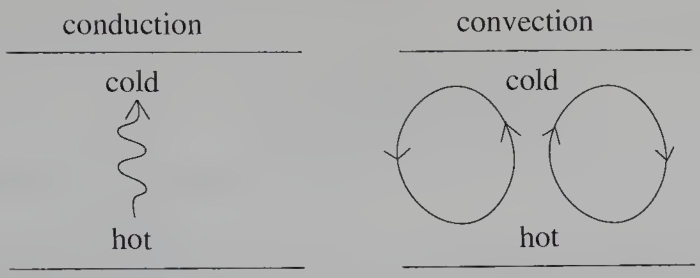

Fig. 12.5 Two alternatives for the heat transport in the Earth. In the left-hand panel the material does not move and heat is transported by conduction. In the right-hand panel the material flows and heat is transported by convection.

This means that ray theory is not valid in or near a caustic. A clear account of the physics of caustics can be found in references [16] and [57]. The former reference contains many beautiful images of caustics.

12.5 Is the Earth's mantle convecting?

The Earth is a body that continuously loses heat to outer space. This heat is partly a remnant of the heat that has been converted from the gravitational energy during the Earth's formation and partly, and more importantly, this heat is generated by the decay of unstable isotopes in the Earth. This heat is transported to the Earth's surface, and the question we aim to address here is: is the heat transported by conduction or by convection?

If the material in the Earth does not flow, heat can only be transported by conduction as shown in the left-hand panel of Figure 12.5. This means that it is the average transfer of the molecular motion from warm regions to cold regions that is responsible for the transport of heat. On the other hand, if the material in the Earth does flow, heat can be carried by the flow as shown in the right-hand panel of Figure 12.5. This process is called convection.

The starting point of the analysis is the heat equation (11.30). In the absence of source terms, this equation, for the special case of a constant heat conduction coefficient κ, may be written as:

$$\frac{\partial T}{\partial t} + \nabla \cdot (\mathbf{v}T) = \kappa \nabla^2 T. \tag{12.32}$$

The term $\nabla \cdot (\mathbf{v}T)$ describes the convective heat transport while the term $\kappa \nabla^2 T$ accounts for the conductive heat transport.

Problem a Let the characteristic velocity be denoted by V, the characteristic length scale by L, and the characteristic temperature perturbation by T. Show that the ratio of the convective heat transport to the conductive heat transport is of the following order:

$$\frac{\text{convective heat transport}}{\text{conductive heat transport}} \sim \frac{VL}{\kappa}. \tag{12.33}$$

This estimate gives the ratio of the two modes of heat transport, but it does not help us much yet because we do not know the order of magnitude V of the flow velocity. This quantity can be obtained from the Navier–Stokes equation of Section 11.6:

$$\frac{\partial(\rho\mathbf{v})}{\partial t} + \nabla\cdot(\rho\mathbf{v}\mathbf{v}) = \mu\nabla^2\mathbf{v} + \mathbf{F}. \tag{11.55}$$

The force $\mathbf{F}$ on the right-hand side is the buoyancy force that is associated with the flow, while the term $\mu\nabla^2\mathbf{v}$ accounts for the viscosity of the flow with viscosity coefficient μ. The mantle of the Earth is extremely viscous and mantle convection (if it exists at all) is a slow process. We therefore assume that the inertia term $\partial(\rho\mathbf{v})/\partial t$ and the advection term $\nabla\cdot(\rho\mathbf{v}\mathbf{v})$ are small compared to the viscous term $\mu\nabla^2\mathbf{v}$. (This assumption would have to be supported by a proper scale analysis.) Under this assumption, the mantle flow is predominantly governed by a balance between the viscous force and the buoyancy force:

$$\mu\nabla^2\mathbf{v} = -\mathbf{F}. \tag{12.34}$$

The next step is to relate the buoyancy to the temperature perturbation T. A temperature perturbation T from a reference temperature T_0 leads to a density perturbation ρ from the reference density ρ_0 given by:

$$\rho = -\alpha T. \tag{12.35}$$

In this expression α is the thermal expansion coefficient which accounts for the expansion or contraction of material due to temperature changes.

Problem b Explain why for most materials $\alpha > 0$. A notable exception is water at temperatures below 4 °C.

Problem c Write $\rho(T_0 + T) = \rho_0 + \rho$ and use the Taylor expansion (3.11) truncated after the first order term to show that the expansion coefficient is given by $\alpha = -\partial\rho/\partial T$.

Problem d The buoyancy force is given by Archimedes's law which states that this force equals the weight of the displaced fluid. Use this result, (12.34), and

(12.35) in a scale analysis to show that the velocity is of the following order:

$$V \sim \frac{g\alpha TL^2}{\mu}, \tag{12.36}$$

where g is the acceleration of gravity.

Problem e Use this to derive that the ratio of the convective heat transport to the conductive heat transport is given by

$$\frac{\text{convective heat transport}}{\text{conductive heat transport}} \sim \frac{g\alpha TL^3}{\mu\kappa}. \tag{12.37}$$

The right-hand side of this expression is dimensionless, and is called the *Rayleigh number* which is denoted by Ra:

$$Ra \equiv \frac{g\alpha TL^3}{\mu\kappa}. \tag{12.38}$$

The Rayleigh number is an indicator for the mode of heat transport. When $Ra \gg 1$, heat is predominantly transported by convection. When the thermal expansion coefficient α is large and when the viscosity μ and the heat conduction coefficient κ are small the Rayleigh number is large and heat is transported by convection.

Problem f Explain physically why a large value of α and small values of μ and κ lead to convective heat transport rather than conductive heat transport.

Dimensionless numbers play a crucial role in fluid mechanics. A discussion of the Rayleigh number and other dimensionless diagnostics such as the Prandtl number and the Grashof number can be found in Section 14.2 of Tritton [108]. The implications of the different values of the Rayleigh number on the character of convection in the Earth's mantle is discussed in references [77] and [110]. Of course, if one wants to use a scale analysis one must know the values of the physical properties involved. For the Earth's mantle, the thermal expansion coefficient α is not very well known because of the complications involved in laboratory measurements of the thermal expansion under the extremely high ambient pressure of the Earth's mantle [29].

12.6 Making an equation dimensionless

Usually the terms in equations that one wants to analyze have a physical dimension such as temperature, velocity, etc. It can sometimes be useful to rescale all the variables in the equation in such a way that the rescaled variables are dimensionless.

This is convenient when setting up numerical solutions of the equations, and it also introduces dimensionless numbers that govern the physics of the problem in a natural way. As an example we will apply this technique to the heat equation (12.32).

Any variable can be made dimensionless by dividing out a constant that has the dimension of the variable. As an example, let the characteristic temperature variation be denoted by T_0, then the dimensional temperature perturbation can be written as:

$$T = T_0 T^{'} . \tag{12.39}$$

The quantity $T^{'}$ is dimensionless. Note that in this section T is the absolute temperature whereas in the previous section T denoted the temperature perturbation.

In this section, dimensionless variables are denoted by a prime. For example, let the characteristic time used to scale the time variable be denoted by τ, then:

$$t = \tau t^{'} . \tag{12.40}$$

We can still leave τ open and later choose a value that simplifies the equations as much as possible. Of course when we want to express the heat equation (12.32) in the new time variable we need to specify how the dimensional time-derivative $\partial/\partial t$ is related to the dimensionless time-derivative $\partial/\partial t^{'}$.

Problem a Use the chain rule for differentiation to show that

$$\frac{\partial}{\partial t} = \frac{1}{\tau}\frac{\partial}{\partial t^{'}} . \tag{12.41}$$

Problem b Let the velocity be scaled with the characteristic velocity (12.36):

$$\mathbf{v} = \frac{g\alpha T_0 L^2}{\mu}\mathbf{v}^{'} , \tag{12.42}$$

and let the position vector be scaled with the characteristic length L of the system: $\mathbf{r} = L\mathbf{r}^{'}$. Use a result similar to (12.41) to convert the spatial derivatives to the new space coordinate and rescale all terms in the heat equation (12.32) to derive the following dimensionless form of this equation

$$\underbrace{\frac{1}{\tau}\frac{\partial T^{'}}{\partial t^{'}}}_{(1)} + \underbrace{\frac{g\alpha T_0 L}{\mu}\nabla^{'} \cdot \left(\mathbf{v}^{'} T^{'}\right)}_{(2)} = \underbrace{\frac{\kappa}{L^2}\nabla^{'2} T}_{(3)} , \tag{12.43}$$

where $\nabla^{'}$ is the gradient operator with respect to the dimensionless coordinates $\mathbf{r}^{'}$.

At this point we have not yet specified the time scale τ for the scaling of the time variable. There are three possible situations that determine τ. First, when term (2) is much larger than term (3), the time derivative (1) balances term (2). Second, when term (3) is much larger than term (2), the time derivative balances term (3). The third possibility is that terms (2) and (3) are of the same order of magnitude, in which case the balance of terms is more complex.

Problem c Show that the ratio of term (2) to term (3) is given by the Rayleigh number which is defined in (12.38).

In this expression all the primed terms are (by definition) of order one. At this point we assume that the convective heat transport dominates the conductive heat transport, that is that $Ra \gg 1$. This means that term (2) is much larger than term (3), hence the time derivative in term (1) must be of the same order as the convective term (2). Terms (1) and (2) can therefore balance only when

$$\frac{1}{\tau} = \frac{g\alpha T_0 L}{\mu}. \tag{12.44}$$

This condition determines the time scale of the evolution of the convecting system.

Problem d Show that with this choice of τ the dimensionless heat equation is given by:

$$\frac{\partial T'}{\partial t'} + \nabla' \cdot \left(\mathbf{v}' T'\right) = \frac{1}{Ra} \nabla'^2 T', \tag{12.45}$$

where Ra is the Rayleigh number.

The advantage of this dimensionless equation over the original heat equation is that (12.45) contains only a single constant Ra, whereas the dimensional heat equation (12.32) depends on a large number of constants. In addition, the scaling of the heat equation has led in a natural way to the key role of the Rayleigh number in the mode of heat transport in a fluid. Since the Rayleigh number is assumed to be large, $1/Ra \ll 1$ and the last term can be seen as a small perturbation. One can either delete this term, or treat it with perturbation theory as described in Chapter 23. The last term in (12.45) contains the highest spatial derivatives of that equation because it contains second order space derivatives whereas the other terms either have no space derivatives (term (1)) or only first order space derivatives (term (2)). Deleting this term turns (12.45) from a second order differential equation (in the space variables) into a first order differential equation. This entails a change in the number of boundary conditions that need to be imposed. As shown in Section 23.7

the last term in (12.45) constitutes a *singular perturbation*, and one should be careful in deleting this (small) term altogether.

Transforming dimensional equations into dimensionless equations is often used to derive the relevant dimensionless physical constants of the system as well as to set up algorithms for solving equations numerically. The basic rationale behind this approach is that the physical units that are used are completely arbitrary. It is immaterial whether we express length in meters or in inches, but of course the numerical value of a given length changes when we change from meters to inches. Making the system dimensionless removes all physical units from the system because all the resulting terms in the equation are dimensionless. This has the drawback that dimensional analysis cannot be used to check for errors, as shown in Section 2.2.

13

Linear algebra

In this chapter several elements of linear algebra are treated that have important applications in physics or that serve to illustrate methodologies used in other areas of mathematical physics.

13.1 Projections and the completeness relation

In mathematical physics, projections play an important role. This is true not only in linear algebra, but also in the analysis of linear systems such as linear filters in data processing (see Section 15.10), and the analysis of vibrating systems such as the normal modes of the Earth which is treated in Section 20.7. Let us consider a vector $\mathbf{v}$ that we want to project along a unit vector $\hat{\mathbf{n}}$, see Figure 13.1. In the examples in this section we work in a three-dimensional space, but the arguments presented here can be generalized to any number of dimensions.

We denote the projection of $\mathbf{v}$ along $\hat{\mathbf{n}}$ as $\mathbf{Pv}$, where $\mathbf{P}$ stands for the projection operator. In a three-dimensional space this operator can be represented by a 3×3 matrix. It is our goal to find the operator $\mathbf{P}$ in terms of the unit vector $\hat{\mathbf{n}}$ as well as the matrix form of this operator. By definition the projection of $\mathbf{v}$ is directed along $\hat{\mathbf{n}}$, hence:

$$\mathbf{Pv} = C\hat{\mathbf{n}}. \tag{13.1}$$

This means that we know the projection operator once the constant C is known.

Problem a Express the length of the vector $\mathbf{Pv}$ in terms of the length of the vector $\mathbf{v}$ and the angle φ of Figure 13.1 and show that $|\mathbf{Pv}| = |\mathbf{v}| \cos\varphi$. Express the angle φ in terms of the inner product of the vectors $\mathbf{v}$ and $\hat{\mathbf{n}}$ by using that $(\hat{\mathbf{n}} \cdot \mathbf{v}) = |\hat{\mathbf{n}}|\,|\mathbf{v}| \cos\varphi = |\mathbf{v}| \cos\varphi$, and use this to show that: $C = (\hat{\mathbf{n}} \cdot \mathbf{v})$.

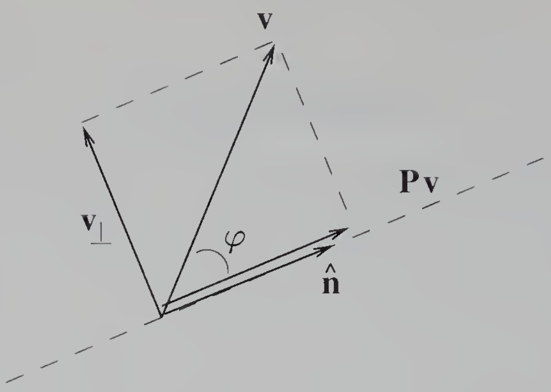

Fig. 13.1 Definition of the geometric variables for the projection of a vector.

Inserting this expression for the constant C in (13.1) leads to an expression for the projection **Pv**:

$$\mathbf{Pv} = \hat{\mathbf{n}}\,(\hat{\mathbf{n}} \cdot \mathbf{v})\,. \tag{13.2}$$

Problem b Show that the component $\mathbf{v}_\perp$ perpendicular to $\hat{\mathbf{n}}$ as defined in Figure 13.1 is given by:

$$\mathbf{v}_\perp = \mathbf{v} - \hat{\mathbf{n}}\,(\hat{\mathbf{n}} \cdot \mathbf{v})\,. \tag{13.3}$$

Problem c As an example, consider the projection along the unit vector along the x-axis: $\hat{\mathbf{n}} = \hat{\mathbf{x}}$. Show using (13.2) and (13.3) that in that case:

$$\mathbf{Pv} = \begin{pmatrix} v_x \\ 0 \\ 0 \end{pmatrix} \quad \text{and} \quad \mathbf{v}_\perp = \begin{pmatrix} 0 \\ v_y \\ v_z \end{pmatrix}.$$

Problem d When we project the vector **Pv** once more along the same unit vector $\hat{\mathbf{n}}$ the vector will not change. We therefore expect that $\mathbf{P}(\mathbf{Pv}) = \mathbf{Pv}$. Show using (13.2) that this is indeed the case. Since this property holds for any vector **v** we can also write it as:

$$\mathbf{P}^2 = \mathbf{P}. \tag{13.4}$$

An operator with this property is called *idempotent*.

Problem e If **P** were a scalar the expression above would imply that **P** is the identity operator **I** or that $\mathbf{P} = 0$. Can you explain why (13.4) does *not* imply that **P** is either the identity operator or equal to zero?

In (13.2) we derived the action of the projection operator on a vector **v**. Since this expression holds for any vector **v** it can be used to derive an explicit form of the projection operator:

$$\mathbf{P} = \hat{\mathbf{n}}\hat{\mathbf{n}}^T. \tag{13.5}$$

This expression should not be confused with the inner product $(\hat{\mathbf{n}} \cdot \hat{\mathbf{n}})$; instead it denotes the dyad of the vector $\hat{\mathbf{n}}$ and itself. The superscript T denotes the transpose of a vector or matrix. The transpose of a vector (or matrix) is found by interchanging its rows and columns. For example, the transpose $\mathbf{A}^T$ of a matrix $\mathbf{A}$ is defined by:

$$A_{ij}^T = A_{ji}, \tag{13.6}$$

and the transpose of the vector $\mathbf{u}$ is defined by:

$$\mathbf{u}^T = (u_x, u_y, u_z) \quad \text{when} \quad \mathbf{u} = \begin{pmatrix} u_x \\ u_y \\ u_z \end{pmatrix}, \tag{13.7}$$

that is taking the transpose converts a column vector into a row vector. The projection operator $\mathbf{P}$ is written in (13.5) as a dyad. In general the dyad $\mathbf{T}$ of two vectors $\mathbf{u}$ and $\mathbf{v}$ is defined as

$$\mathbf{T} = \mathbf{u}\mathbf{v}^T. \tag{13.8}$$

This expression simply means that the components T_{ij} of the dyad are defined by

$$T_{ij} = u_i v_j, \tag{13.9}$$

where u_i is the i-component of $\mathbf{u}$ and v_j is the j-component of $\mathbf{v}$.

In the literature you will find different notations for the inner product of two vectors. The inner product of the vectors $\mathbf{u}$ and $\mathbf{v}$ is sometimes written as

$$(\mathbf{u} \cdot \mathbf{v}) = \mathbf{u}^T\mathbf{v}. \tag{13.10}$$

Problem f Considering the vector $\mathbf{v}$ as a 3×1 matrix and the vector $\mathbf{v}^T$ as a 1×3 matrix, show that the notation used on the right-hand sides of (13.10) and (13.8) is consistent with the normal rules for matrix multiplication.

Equation (13.5) relates the projection operator $\mathbf{P}$ to the unit vector $\hat{\mathbf{n}}$. From this the representation of the projection operator as a 3×3 matrix can be found by computing the dyad $\hat{\mathbf{n}}\hat{\mathbf{n}}^T$.

Problem g Show that the operator for the projection along the unit vector

$$\hat{\mathbf{n}} = \frac{1}{\sqrt{14}} \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}$$

is given by

$$\mathbf{P} = \frac{1}{14}\begin{pmatrix} 1 & 2 & 3 \\ 2 & 4 & 6 \\ 3 & 6 & 9 \end{pmatrix}.$$

Verify explicitly that for this example $\mathbf{P}\hat{\mathbf{n}} = \hat{\mathbf{n}}$, and explain this result.

Up to this point we have projected the vector $\mathbf{v}$ along a single unit vector $\hat{\mathbf{n}}$. Suppose we have a set of mutually orthogonal unit vectors $\hat{\mathbf{n}}_i$. The fact that these unit vectors are mutually orthogonal means that the different unit vectors are perpendicular to each other: $\left(\hat{\mathbf{n}}_i \cdot \hat{\mathbf{n}}_j\right) = 0$ when $i \neq j$. We can project $\mathbf{v}$ on each of these unit vectors and add these projections. This gives us the projection of $\mathbf{v}$ on the subspace spanned by the unit vectors $\hat{\mathbf{n}}_i$:

$$\mathbf{P}\mathbf{v} = \sum_i \hat{\mathbf{n}}_i \left(\hat{\mathbf{n}}_i \cdot \mathbf{v}\right). \tag{13.11}$$

When the unit vectors $\hat{\mathbf{n}}_i$ span the full space we are working in, the projected vector is identical to the original vector. To see this, consider for example a three-dimensional space. Any vector can be decomposed into its components along the x-, y-, and z-axes, and this can be written as:

$$\mathbf{v} = v_x\hat{\mathbf{x}} + v_y\hat{\mathbf{y}} + v_z\hat{\mathbf{z}} = \hat{\mathbf{x}}\left(\hat{\mathbf{x}} \cdot \mathbf{v}\right) + \hat{\mathbf{y}}\left(\hat{\mathbf{y}} \cdot \mathbf{v}\right) + \hat{\mathbf{z}}\left(\hat{\mathbf{z}} \cdot \mathbf{v}\right); \tag{13.12}$$

note that this expression has the same form as (13.11). This implies that when in (13.11) we sum over a set of unit vectors that completely spans the space we are working in, the right-hand side of (13.11) is identical to the original vector $\mathbf{v}$, that is $\sum_i \hat{\mathbf{n}}_i \left(\hat{\mathbf{n}}_i \cdot \mathbf{v}\right) = \mathbf{v}$. The operator of the left-hand side of this equality is therefore identical to the identity operator $\mathbf{I}$:

$$\sum_{i=1}^{N} \hat{\mathbf{n}}_i \hat{\mathbf{n}}_i^T = \mathbf{I}. \tag{13.13}$$

The dimension of the space we are working in is N; if we sum over a smaller number of unit vectors we project on a subspace of the N-dimensional space. Expression (13.13) expresses that the vectors $\hat{\mathbf{n}}_i$ (with $i = 1, \ldots, N$) can be used to give a complete representation of any vector. Such a set of vectors is called a *complete set*, and expression (13.13) is called the *closure relation*.

Problem h Verify explicitly that when the unit vectors $\hat{\mathbf{n}}_i$ are chosen to be the unit vectors $\hat{\mathbf{x}}$, $\hat{\mathbf{y}}$, and $\hat{\mathbf{z}}$ along the x-, y-, and z-axes that the right-hand side of (13.13) is given by the 3×3 identity matrix.

There are, of course, many different ways of choosing a set of three orthogonal unit vectors in three dimensions. Expression (13.13) should hold for every choice of a complete set of unit vectors.

Problem i Verify explicitly that when the unit vectors $\hat{\mathbf{n}}_i$ are chosen to be the unit vectors $\hat{\mathbf{r}}$, $\hat{\boldsymbol{\theta}}$, and $\hat{\boldsymbol{\varphi}}$ defined in (4.6) for a system of spherical coordinates the right-hand side of (13.13) is given by the 3 × 3 identity matrix.

13.2 Projection on vectors that are not orthogonal

In the previous section we considered a projection on a set of orthogonal unit vectors. In this section we consider an example of a projection on a set of vectors that are not necessarily orthogonal. Consider two vectors **a** and **b** in a three-dimensional space. These two vectors span a two-dimensional plane. In this section we determine the projection of a vector **v** on the plane spanned by the vectors **a** and **b**, see Figure 13.2 for the geometry of the problem. The projection of **v** on the plane is denoted by **Pv**.

By definition the projected vector **Pv** lies in the plane spanned by **a** and **b**, and this vector can thus be written as:

$$\mathbf{Pv} = \alpha\mathbf{a} + \beta\mathbf{b}. \tag{13.14}$$

The task of finding the projection can therefore be reduced to finding the two coefficients α and β. These constants follow from the requirement that the vector joining **v** with its projection **Pv** is perpendicular to both **a** and **b**, see Figure 13.2. This means that $(\mathbf{v} - \mathbf{Pv}) \cdot \mathbf{a} = (\mathbf{v} - \mathbf{Pv}) \cdot \mathbf{b} = 0$.

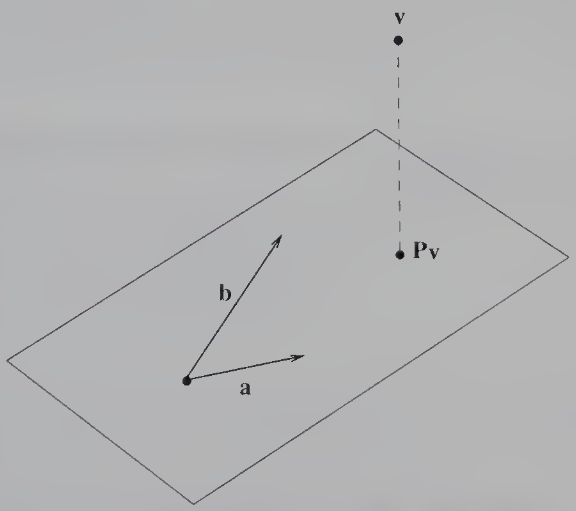

Fig. 13.2 Definition of the geometric variables for the projection on a plane.

Problem a Show that this requirement is equivalent to the following system of equations for α and β:

$$\left.\begin{aligned}\alpha\,(\mathbf{a}\cdot\mathbf{a})+\beta\,(\mathbf{a}\cdot\mathbf{b})&=(\mathbf{a}\cdot\mathbf{v})\,,\\ \alpha\,(\mathbf{a}\cdot\mathbf{b})+\beta\,(\mathbf{b}\cdot\mathbf{b})&=(\mathbf{b}\cdot\mathbf{v})\,.\end{aligned}\right\} \tag{13.15}$$

Problem b Show that the solution of this system is given by

$$\left.\begin{aligned}\alpha&=\frac{\left[b^2\mathbf{a}-(\mathbf{a}\cdot\mathbf{b})\,\mathbf{b}\right]\cdot\mathbf{v}}{a^2b^2-(\mathbf{a}\cdot\mathbf{b})^2},\\ \beta&=\frac{\left[a^2\mathbf{b}-(\mathbf{a}\cdot\mathbf{b})\,\mathbf{a}\right]\cdot\mathbf{v}}{a^2b^2-(\mathbf{a}\cdot\mathbf{b})^2},\end{aligned}\right\} \tag{13.16}$$

where a denotes the length of the vector **a**: $a\equiv|\mathbf{a}|$, and a similar notation is used for the vector **b**.

Problem c Show using (13.14) and (13.16) that the projection operator **P** for the projection on the plane is given by

$$\mathbf{P}=\frac{1}{a^2b^2-(\mathbf{a}\cdot\mathbf{b})^2}\left[b^2\mathbf{a}\mathbf{a}^T+a^2\mathbf{b}\mathbf{b}^T-(\mathbf{a}\cdot\mathbf{b})\left(\mathbf{a}\mathbf{b}^T+\mathbf{b}\mathbf{a}^T\right)\right]. \tag{13.17}$$

This example shows that a projection on a set of nonorthogonal basis vectors is much more complex than the projection on a set of orthonormal basis vectors. A different way of finding the projection operator of expression (13.17) is by first finding two *orthogonal* unit vectors in the plane spanned by **a** and **b** and then using (13.11). One unit vector can be found by dividing **a** by its length to give the unit vector $\hat{\mathbf{a}}=\mathbf{a}/|\mathbf{a}|$. The second unit vector can be found by considering the component $\mathbf{b}_\perp$ of **b** perpendicular to $\hat{\mathbf{a}}$ and by normalizing the resulting vector to form the unit vector $\hat{\mathbf{b}}_\perp$ that is perpendicular to $\hat{\mathbf{a}}$, see Figure 13.3.

Problem d Use (13.3) to find $\hat{\mathbf{b}}_\perp$ and show that the projection operator **P** of (13.17) can also be written as

$$\mathbf{P}=\hat{\mathbf{a}}\hat{\mathbf{a}}^T+\hat{\mathbf{b}}_\perp\hat{\mathbf{b}}_\perp^T. \tag{13.18}$$

Note that this expression is consistent with (13.11).

Up to this point the plane has been defined by the vectors **a** and **b** (or equivalently by the orthonormal unit vectors $\hat{\mathbf{a}}$ and $\hat{\mathbf{b}}_\perp$). However, a plane can also be defined by the unit vector $\hat{\mathbf{n}}$ that is perpendicular to the plane, see Figure 13.3. In fact, the unit vectors $\hat{\mathbf{a}}$, $\hat{\mathbf{b}}_\perp$ and $\hat{\mathbf{n}}$ form a complete orthonormal basis of the three-dimensional

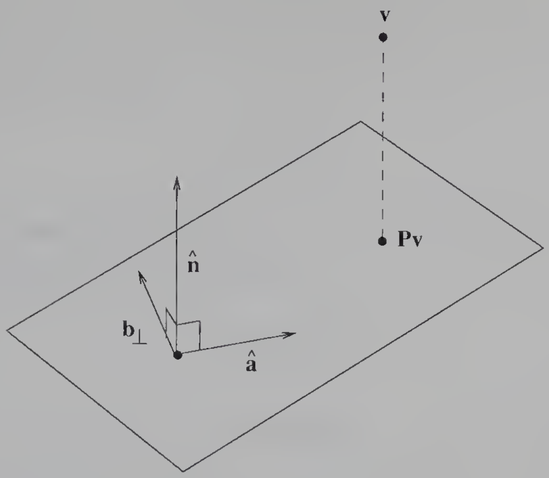

Fig. 13.3 Definition of the normal vector to a plane.

space. According to (13.13) this implies that $\hat{\mathbf{a}}\hat{\mathbf{a}}^T + \hat{\mathbf{b}}_\perp\hat{\mathbf{b}}_\perp^T + \hat{\mathbf{n}}\hat{\mathbf{n}}^T = \mathbf{I}$. With (13.18) this implies that the projection operator $\mathbf{P}$ can also be written as

$$\mathbf{P} = \mathbf{I} - \hat{\mathbf{n}}\hat{\mathbf{n}}^T . \tag{13.19}$$

Problem e Give an alternative derivation of this result. Hint: let the operator in (13.19) act on an arbitrary vector $\mathbf{v}$.

13.3 Coriolis force and centrifugal force

As an example of working with the cross-product of vectors we consider the inertia forces that occur in the mechanics of rotating coordinate systems. This is of importance in the Earth sciences, because the rotation of the Earth plays a crucial role in the motion of wind in the atmosphere and currents in the ocean. In addition, the Earth's rotation is essential for the generation of the magnetic field of the Earth in the outer core.

In order to describe the motion of a particle in a rotating coordinate system we need to characterize the rotation somehow. This can be achieved by introducing a vector $\mathbf{\Omega}$ that is aligned with the rotation axis and whose length is given by the *rate* of rotation expressed in radians per seconds.

Problem a What is the direction of $\mathbf{\Omega}$ and the length $\Omega = |\mathbf{\Omega}|$ for the Earth's rotation.

Let us assume we are considering a vector $\mathbf{q}$ that is constant in the rotating coordinate system. In a nonrotating system this vector changes with time because it co-rotates

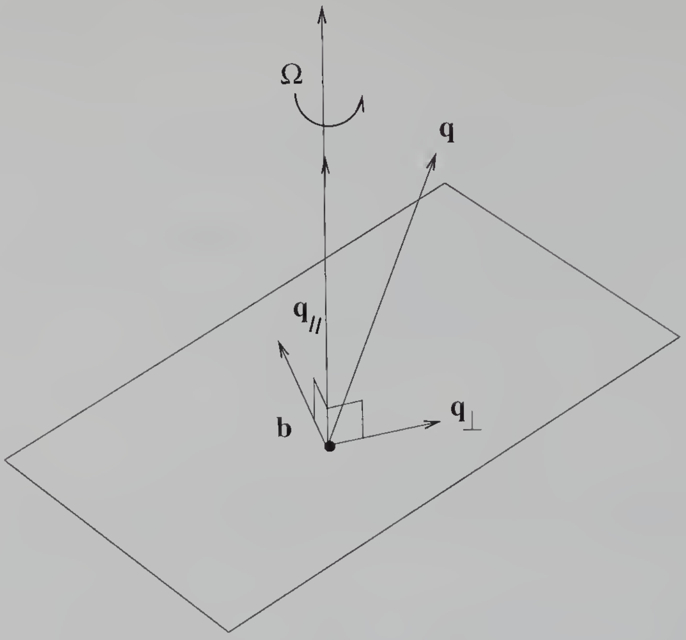

Fig. 13.4 Decomposition of a vector in a rotating coordinate system.

with the rotating system. The vector $\mathbf{q}$ can be decomposed into a component $\mathbf{q}_{//}$ along the rotation vector and a component $\mathbf{q}_\perp$ perpendicular to the rotation vector. In addition, a vector $\mathbf{b}$ is defined in Figure 13.4 that is perpendicular to both $\mathbf{q}_\perp$ and $\boldsymbol{\Omega}$ in such a way that $\boldsymbol{\Omega}$, $\mathbf{q}_\perp$ and $\mathbf{b}$ form a right-handed orthogonal system.

Problem b Show that:

$$\left.\begin{aligned} \mathbf{q}_{//} &= \hat{\boldsymbol{\Omega}}\left(\hat{\boldsymbol{\Omega}}\cdot\mathbf{q}\right), \\ \mathbf{q}_\perp &= \mathbf{q}-\hat{\boldsymbol{\Omega}}\left(\hat{\boldsymbol{\Omega}}\cdot\mathbf{q}\right), \\ \mathbf{b} &= \hat{\boldsymbol{\Omega}}\times\mathbf{q}. \end{aligned}\right\} \tag{13.20}$$

The last identity follows by inserting the second equation into the identity $\mathbf{b}=\hat{\boldsymbol{\Omega}}\times\mathbf{q}_\perp$.

Problem c In a fixed nonrotating coordinate system, the vector $\mathbf{q}$ rotates, hence its position is time-dependent: $\mathbf{q}=\mathbf{q}(t)$. Let us consider how the vector changes over a time interval Δt. Since the component $\mathbf{q}_{//}$ is at all times directed along the rotation vector $\boldsymbol{\Omega}$, it is constant in time. Over a time interval Δt the coordinate system rotates over an angle $\Omega\Delta t$. Use this to show that the component of $\mathbf{q}$ perpendicular to the rotation vector satisfies:

$$\mathbf{q}_\perp(t+\Delta t)=\cos(\Omega\Delta t)\,\mathbf{q}_\perp(t)+\sin(\Omega\Delta t)\,\mathbf{b}, \tag{13.21}$$

and that time evolution of $\mathbf{q}$ is therefore given by

$$\mathbf{q}(t+\Delta t)=\mathbf{q}(t)+[\cos(\Omega\Delta t)-1]\,\mathbf{q}_\perp(t)+\sin(\Omega\Delta t)\,\mathbf{b}. \tag{13.22}$$

Problem d The goal is to obtain the time derivative of the vector $\mathbf{q}$. This quantity can be computed using the rule $d\mathbf{q}/dt = \lim_{\Delta t \to 0}[\mathbf{q}(t + \Delta t) - \mathbf{q}(t)]/\Delta t$. Use this and (13.22) to show that

$$\dot{\mathbf{q}} = \Omega \mathbf{b}, \tag{13.23}$$

where the dot denotes the time derivative. Use (13.20) to show that the time derivative of the vector $\mathbf{q}$ is given by

$$\dot{\mathbf{q}} = \boldsymbol{\Omega} \times \mathbf{q}. \tag{13.24}$$

At this point the vector $\mathbf{q}$ can be any vector that co-rotates with the rotating coordinate system. In this rotating coordinate system, three Cartesian basis vectors $\hat{\mathbf{x}}$, $\hat{\mathbf{y}}$, and $\hat{\mathbf{z}}$ can be used as a basis to decompose the position vector:

$$\mathbf{r}_{rot} = x\hat{\mathbf{x}} + y\hat{\mathbf{y}} + z\hat{\mathbf{z}}. \tag{13.25}$$

Since these basis vectors are constant in the rotating coordinate system, they satisfy (13.24) so that:

$$\left.\begin{aligned} d\hat{\mathbf{x}}/dt &= \boldsymbol{\Omega} \times \hat{x}, \\ d\hat{\mathbf{y}}/dt &= \boldsymbol{\Omega} \times \hat{y}, \\ d\hat{\mathbf{z}}/dt &= \boldsymbol{\Omega} \times \hat{z}. \end{aligned}\right\} \tag{13.26}$$

It should be noted that we have *not* assumed that the position vector $\mathbf{r}_{rot}$ in (13.25) rotates with the coordinate system, we have only assumed that the unit vectors $\hat{\mathbf{x}}$, $\hat{\mathbf{y}}$, and $\hat{\mathbf{z}}$ rotate with the coordinate system. Of course, this will leave an imprint on the velocity and the acceleration. In general the velocity and the acceleration follow by differentiating (13.25) with time. If the unit vectors $\hat{\mathbf{x}}$, $\hat{\mathbf{y}}$, and $\hat{\mathbf{z}}$ were fixed, they would not contribute to the time derivative. However, the unit vectors $\hat{\mathbf{x}}$, $\hat{\mathbf{y}}$, and $\hat{\mathbf{z}}$ rotate with the coordinate system and the associated time derivative is given by (13.26).

Problem e Differentiate the position vector in (13.25) with respect to time and show that the velocity vector $\mathbf{v}$ is given by:

$$\mathbf{v} = \dot{x}\hat{\mathbf{x}} + \dot{y}\hat{\mathbf{y}} + \dot{z}\hat{\mathbf{z}} + \boldsymbol{\Omega} \times \mathbf{r}_{rot}. \tag{13.27}$$

The terms $\dot{x}\hat{\mathbf{x}} + \dot{y}\hat{\mathbf{y}} + \dot{z}\hat{\mathbf{z}}$ denote the velocity as seen in the rotating coordinate system; this velocity is denoted by $\mathbf{v}_{rot}$. The velocity vector can therefore be written as:

$$\mathbf{v} = \mathbf{v}_{rot} + \boldsymbol{\Omega} \times \mathbf{r}_{rot}. \tag{13.28}$$

Problem f Give an interpretation of the last term in this expression.

Problem g The acceleration follows by differentiating (13.27) for the velocity once more with respect to time. Show that the acceleration is given by

$$\mathbf{a} = \ddot{x}\hat{\mathbf{x}} + \ddot{y}\hat{\mathbf{y}} + \ddot{z}\hat{\mathbf{z}} + 2\boldsymbol{\Omega} \times (\dot{x}\hat{\mathbf{x}} + \dot{y}\hat{\mathbf{y}} + \dot{z}\hat{\mathbf{z}}) + \boldsymbol{\Omega} \times (\boldsymbol{\Omega} \times \mathbf{r}_{rot}) . \tag{13.29}$$

The terms $\ddot{x}\hat{\mathbf{x}} + \ddot{y}\hat{\mathbf{y}} + \ddot{z}\hat{\mathbf{z}}$ on the right-hand side denote the acceleration as seen in the rotating coordinate system; this quantity will be denoted by $\mathbf{a}_{rot}$. The terms $\dot{x}\hat{\mathbf{x}} + \dot{y}\hat{\mathbf{y}} + \dot{z}\hat{\mathbf{z}}$ again denote the velocity $\mathbf{v}_{rot}$ as seen in the rotating coordinate system. The left-hand side is by Newton's law equal to $\mathbf{F}/m$, where $\mathbf{F}$ is the force acting on the particle.

Problem h Use this to show that in the rotating coordinate system Newton's law is given by:

$$m\mathbf{a}_{rot} = \mathbf{F} - 2m\boldsymbol{\Omega} \times v_{rot} - m\boldsymbol{\Omega} \times (\boldsymbol{\Omega} \times \mathbf{r}_{rot}) . \tag{13.30}$$

The rotation manifests itself through two additional forces. The term $-2m\boldsymbol{\Omega} \times v_{rot}$ describes the *Coriolis force* and the term $-m\boldsymbol{\Omega} \times (\boldsymbol{\Omega} \times \mathbf{r}_{rot})$ describes the centrifugal force.

Problem i Show that the centrifugal force is perpendicular to the rotation axis and is directed from the rotation axis towards the particle.

Problem j Air flows from high-pressure areas to low-pressure areas. As air flows in the northern hemisphere from a high-pressure area to a low-pressure area, is it deflected towards the right or towards the left when seen from above?

Problem k Compute the magnitude of the centrifugal force and the Coriolis force you experience due to the Earth's rotation when you ride your bicycle. Compare this with the force mg you experience due to the gravitational attraction of the Earth. It suffices to compute the orders of magnitude of the different terms. In the northern hemisphere does the Coriolis force deflect you to the left or to the right? Have you ever noticed a tilt due to the Coriolis force while riding your bicycle?

In meteorology and oceanography it is often convenient to describe the motion of air or water along the Earth's surface using a Cartesian coordinate system that rotates with the Earth with unit vectors pointing eastwards ($\hat{\mathbf{e}}_1$), northwards ($\hat{\mathbf{e}}_2$), and upwards ($\hat{\mathbf{e}}_3$), see Figure 13.5. These unit vectors can be related to the unit vectors

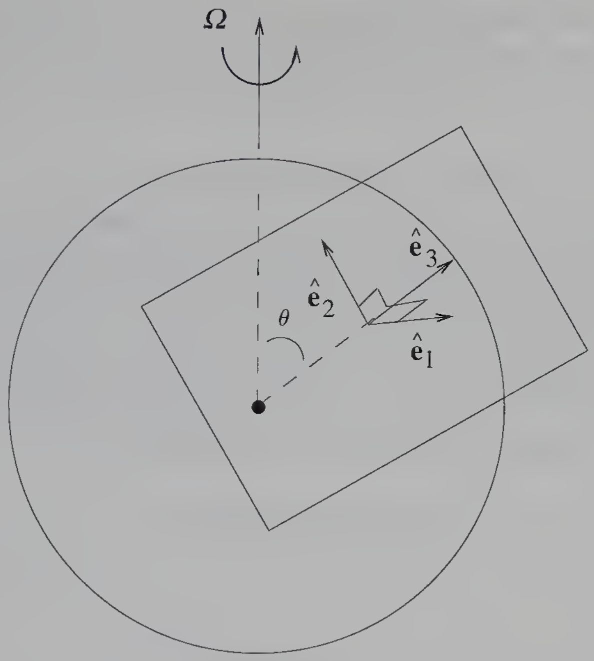

Fig. 13.5 Definition of a local Cartesian coordinate system that is aligned with the Earth's surface.

$\hat{\mathbf{r}}$, $\hat{\boldsymbol{\varphi}}$ and $\hat{\boldsymbol{\theta}}$ that are defined in (4.7). Let the velocity in the eastward direction be denoted by u, the velocity in the northward direction by v, and the vertical velocity by w.

Problem l Show that:

$$\hat{\mathbf{e}}_1 = \hat{\boldsymbol{\varphi}}, \quad \hat{\mathbf{e}}_2 = -\hat{\boldsymbol{\theta}}, \quad \hat{\mathbf{e}}_3 = \hat{\mathbf{r}}, \tag{13.31}$$

and that the velocity in this rotating coordinate system is given by

$$\mathbf{v} = u\hat{\mathbf{e}}_1 + v\hat{\mathbf{e}}_2 + w\hat{\mathbf{e}}_3 . \tag{13.32}$$

Problem m We assume that the axes of the spherical coordinate system are chosen in such a way that the direction $\theta = 0$ is aligned with the rotation axis. This is a different way of saying that the rotation vector is parallel to the z-axis: $\boldsymbol{\Omega} = \Omega\hat{\mathbf{z}}$. Use the first two expressions of (4.13) to show that the rotation vector has the following expansion in terms of the unit vectors $\hat{\mathbf{r}}$ and $\hat{\boldsymbol{\theta}}$:

$$\boldsymbol{\Omega} = \Omega\left(\cos\theta\ \hat{\mathbf{r}} - \sin\theta\ \hat{\boldsymbol{\theta}}\right). \tag{13.33}$$

Problem n In the rotating coordinate system, the Coriolis force is given by $\mathbf{F}_{cor} = -2m\boldsymbol{\Omega} \times v$. Use expressions (13.31)–(13.33) and (4.11) for the cross-product of the unit vectors to show that the Coriolis force is given by

$$\mathbf{F}_{cor} = 2m\Omega \sin\theta \; u \; \hat{\mathbf{r}} + 2m\Omega \cos\theta \; u \; \hat{\boldsymbol{\theta}} + 2m\Omega (v \cos\theta \; - w \sin\theta)\, \hat{\boldsymbol{\varphi}}. \quad (13.34)$$

Problem o Both the ocean and the atmosphere are shallow in the sense that the vertical length scale (a few kilometers for the ocean and around 10 kilometers for the atmosphere) is much less than the horizontal length scale. This causes the vertical velocity to be much smaller than the horizontal velocity. For this reason the vertical velocity w can be neglected in (13.34). Use this approximation and definition (13.31) to show that the horizontal component $\mathbf{a}_{cor}^{H}$ of the Coriolis acceleration is given in this approach by:

$$\mathbf{a}_{cor}^{H} = -f\hat{\mathbf{e}}_3 \times \mathbf{v}, \quad (13.35)$$

with

$$f = 2\Omega \cos\theta. \quad (13.36)$$

This result is widely used in meteorology and oceanography, because (13.35) states that, in the Cartesian coordinate system aligned with the Earth's surface, the Coriolis force generated by the rotation around the true Earth's axis of rotation is identical to the Coriolis force generated by the rotation around a vertical axis with a rotation rate given by $\Omega \cos\theta$. This rotation rate is largest at the poles, where $\cos\theta = \pm 1$, and vanishes at the equator, where $\cos\theta = 0$. The parameter f in (13.35) acts as a coupling parameter; it is called the *Coriolis parameter.* (In the literature on geophysical fluid dynamics one often uses latitude rather than the colatitude θ that is used here, and for this reason one often sees a sin term rather than a cos term in the definition of the Coriolis parameter.) In many applications one disregards the dependence of f on the colatitude θ; in that approach f is a constant and one speaks of the *f-plane approximation.* However, the dependence of the Coriolis parameter on θ is crucial in explaining a number of atmospheric and oceanographic phenomena such as the propagation of Rossby waves and the formation of the Gulfstream. In a further refinement one linearizes the dependence of the Coriolis parameter with colatitude. This leads to the *β-plane approximation.* Details can be found in the books of Holton [50] and Pedlosky [82].

13.4 Eigenvalue decomposition of a square matrix

In this section we consider the way in which a square $N \times N$ matrix $\mathbf{A}$ operates on a vector. Since a matrix describes a linear transformation from a vector to a new vector, the action of the matrix $\mathbf{A}$ can be quite complicated. However, suppose the matrix has a set of eigenvectors $\hat{\mathbf{v}}^{(n)}$. We assume these eigenvectors are normalized, hence a caret is used in the notation $\hat{\mathbf{v}}^{(n)}$. These eigenvectors are useful because the action of $\mathbf{A}$ on an eigenvector $\hat{\mathbf{v}}^{(n)}$ is simple:

$$\mathbf{A}\hat{\mathbf{v}}^{(n)} = \lambda_n \hat{\mathbf{v}}^{(n)}, \tag{13.37}$$

where λ_n is the eigenvalue of the eigenvector $\hat{\mathbf{v}}^{(n)}$. When $\mathbf{A}$ acts on an eigenvector, the resulting vector is parallel to the original vector, and the only effect of $\mathbf{A}$ on this vector is to elongate the vector (when $\lambda_n \geq 1$), compress the vector (when $0 \leq \lambda_n < 1$), or reverse the vector (when $\lambda_n < 0$). We restrict ourselves here to matrices that are real and symmetric.

Problem a Show that for such a matrix the eigenvalues are real and the eigenvectors are orthogonal.

The fact that the eigenvectors $\hat{\mathbf{v}}^{(n)}$ are normalized and mutually orthogonal can be expressed as

$$\left(\hat{\mathbf{v}}^{(n)} \cdot \hat{\mathbf{v}}^{(m)}\right) = \delta_{nm}, \tag{13.38}$$

where δ_{nm} is the Kronecker delta which is defined as

$$\delta_{nm} = \begin{cases} 1 & \text{when} \quad n = m \\ 0 & \text{when} \quad n \neq m \end{cases}. \tag{11.15}$$

The eigenvectors $\hat{\mathbf{v}}^{(n)}$ can be used to define the columns of a matrix $\mathbf{V}$:

$$\mathbf{V} = \begin{pmatrix} \vdots & \vdots & & \vdots \\ \hat{\mathbf{v}}^{(1)} & \hat{\mathbf{v}}^{(2)} & \cdots & \hat{\mathbf{v}}^{(N)} \\ \vdots & \vdots & & \vdots \end{pmatrix}; \tag{13.39}$$

this definition implies that

$$V_{ij} \equiv v_i^{(j)}. \tag{13.40}$$

Problem b Use the orthogonality of the eigenvectors $\hat{\mathbf{v}}^{(n)}$ (expression (13.38)) to show that the matrix $\mathbf{V}$ is unitary, that is to show that

$$\mathbf{V}^T\mathbf{V} = \mathbf{I}, \tag{13.41}$$

where $\mathbf{I}$ is the identity matrix with elements $I_{kl} = \delta_{kl}$. The superscript T denotes the transpose.

There are N eigenvectors which are mutually orthonormal in an N-dimensional space. These eigenvectors therefore form a complete set and analogously to (13.13) the completeness relation can be expressed as

$$\mathbf{I} = \sum_{n=1}^{N} \hat{\mathbf{v}}^{(n)}\hat{\mathbf{v}}^{(n)T} . \tag{13.42}$$

When the terms in this expression operate on an arbitrary vector $\mathbf{p}$, an expansion of $\mathbf{p}$ in the eigenvectors that is analogous to (13.11) is obtained:

$$\mathbf{p} = \sum_{n=1}^{N} \hat{\mathbf{v}}^{(n)}\hat{\mathbf{v}}^{(n)T}\mathbf{p} = \sum_{n=1}^{N} \hat{\mathbf{v}}^{(n)}\left(\hat{\mathbf{v}}^{(n)} \cdot \mathbf{p}\right) . \tag{13.43}$$

This is a useful expression because it can be used to simplify the effect of the matrix $\mathbf{A}$ on an arbitrary vector $\mathbf{p}$.

Problem c Let $\mathbf{A}$ act on (13.43) and show that:

$$\mathbf{A}\mathbf{p} = \sum_{n=1}^{N} \lambda_n \hat{\mathbf{v}}^{(n)}\left(\hat{\mathbf{v}}^{(n)} \cdot \mathbf{p}\right) . \tag{13.44}$$

This expression has an interesting geometric interpretation. When $\mathbf{A}$ acts on $\mathbf{p}$, the vector $\mathbf{p}$ is projected on each of the eigenvectors; this is described by the term $(\hat{\mathbf{v}}^{(n)} \cdot \mathbf{p})$. The corresponding eigenvector $\hat{\mathbf{v}}^{(n)}$ is multiplied by the eigenvalue $\hat{\mathbf{v}}^{(n)} \rightarrow \lambda_n \hat{\mathbf{v}}^{(n)}$, and the result is summed over all the eigenvectors. The action of $\mathbf{A}$ can thus be reduced to a projection on eigenvectors, a multiplication with the corresponding eigenvalue and a summation over all eigenvectors. The eigenvalue λ_n can be seen as the sensitivity of the eigenvector $\hat{\mathbf{v}}^{(n)}$ to the matrix $\mathbf{A}$.

Problem d Expression (13.44) holds for every vector $\mathbf{p}$. Use this to show that $\mathbf{A}$ can be written as:

$$\mathbf{A} = \sum_{n=1}^{N} \lambda_n \hat{\mathbf{v}}^{(n)}\hat{\mathbf{v}}^{(n)T} . \tag{13.45}$$

Problem e Show that with the definition (13.39) this result can also be written as:

$$\mathbf{A} = \mathbf{V}\Sigma\mathbf{V}^T , \tag{13.46}$$

where Σ is a matrix that has the eigenvalues on its diagonal and whose other elements are equal to zero:

$$\Sigma = \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & & \lambda_N \end{pmatrix}. \tag{13.47}$$

Hint: let (13.46) act on an arbitrary vector, use definition (13.40) and see what happens.

13.5 Computing a function of a matrix

Expansion (13.45) (or equivalently (13.46)) is useful because it provides a way to compute the inverse of a matrix and to compute functions of a matrix such as the exponential of a matrix. Let us first use (13.45) to compute the inverse $\mathbf{A}^{-1}$ of the matrix. In order to do this we must know the effect of $\mathbf{A}^{-1}$ on the eigenvectors $\hat{\mathbf{v}}^{(n)}$.

Problem a Use the identities $\hat{\mathbf{v}}^{(n)} = \mathbf{I}\hat{\mathbf{v}}^{(\mathbf{n})} = \mathbf{A}^{-1}\mathbf{A}\hat{\mathbf{v}}^{(n)} = \mathbf{A}^{-1}\lambda_n\hat{\mathbf{v}}^{(n)}$ to show that $\hat{\mathbf{v}}^{(n)}$ is also an eigenvector of the inverse $\mathbf{A}^{-1}$ with eigenvalue $1/\lambda_n$:

$$\mathbf{A}^{-1}\hat{\mathbf{v}}^{(n)} = \frac{1}{\lambda_n}\hat{\mathbf{v}}^{(n)}. \tag{13.48}$$

Problem b Use this result and the eigenvector decomposition (13.43) to show that the effect of $\mathbf{A}^{-1}$ on a vector $\mathbf{p}$ can be written as

$$\mathbf{A}^{-1}\mathbf{p} = \sum_{n=1}^{N} \frac{1}{\lambda_n}\hat{\mathbf{v}}^{(n)}\left(\hat{\mathbf{v}}^{(n)} \cdot \mathbf{p}\right). \tag{13.49}$$

Also show that this implies that $\mathbf{A}^{-1}$ can also be written as:

$$\mathbf{A}^{-1} = \mathbf{V}\Sigma^{-1}\mathbf{V}^T, \tag{13.50}$$

with

$$\Sigma^{-1} = \begin{pmatrix} 1/\lambda_1 & 0 & \cdots & 0 \\ 0 & 1/\lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & & 1/\lambda_N \end{pmatrix}. \tag{13.51}$$

This is an important result: it means that once we have computed the eigenvectors and eigenvalues of a matrix, we can compute the inverse matrix efficiently. Note that this procedure gives problems when one of the eigenvalues is equal to zero because for such an eigenvalue $1/\lambda_n$ is not defined. This makes sense: when one (or more) of the eigenvalues vanishes, the matrix is singular and the inverse does not exist. Also when one of the eigenvalues is nonzero but close to zero, the corresponding term $1/\lambda_n$ is large, and in practice this gives rise to numerical instabilities. In this situation the inverse of the matrix exists, but the result is very sensitive to computational (and other) errors. Such a matrix is called *poorly conditioned.*

In general, a function of a matrix, such as the exponent of a matrix, is not defined. However, suppose we have a function $f(z)$ that operates on a scalar z and that this function can be written as a power series:

$$f(z) = \sum_p a_p z^p . \tag{13.52}$$

For example, when $f(z) = \exp(z)$, then $f(z) = \sum_{p=0}^{\infty}(1/p!)z^p$. On replacing the scalar z by the matrix $\mathbf{A}$ the power series expansion can be used to *define* the effect of the function f when it operates on the matrix $\mathbf{A}$:

$$f(\mathbf{A}) \equiv \sum_p a_p \mathbf{A}^p . \tag{13.53}$$

Although this may seem to be a simple rule for computing $f(\mathbf{A})$, it is actually not so useful because in many applications the summation (13.53) consists of infinitely many terms and the computation of $\mathbf{A}^p$ can be computationally very demanding. Again, the eigenvalue decomposition (13.45) or (13.46) allows us to simplify the evaluation of $f(\mathbf{A})$.

Problem c Show that $\hat{\mathbf{v}}^{(n)}$ is also an eigenvector of $\mathbf{A}^p$ with eigenvalue $(\lambda_n)^p$, that is show that

$$\mathbf{A}^p \hat{\mathbf{v}}^{(n)} = (\lambda_n)^p \, \hat{\mathbf{v}}^{(n)} . \tag{13.54}$$

Hint: first compute $\mathbf{A}^2\hat{\mathbf{v}}^{(n)} = \mathbf{A}\left(\mathbf{A}\hat{\mathbf{v}}^{(n)}\right)$, then $\mathbf{A}^3\hat{\mathbf{v}}^{(n)}$, etc.

Problem d Use this result to show that (13.46) can be generalized to:

$$\mathbf{A}^p = \mathbf{V}\Sigma^p\mathbf{V}^T , \tag{13.55}$$

with Σ^p given by

$$\Sigma^p = \begin{pmatrix} \lambda_1^p & 0 & \cdots & 0 \\ 0 & \lambda_2^p & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & & \lambda_N^p \end{pmatrix} . \tag{13.56}$$

Problem e Finally use (13.52) and (13.53) to show that $f(\mathbf{A})$ can be written as:

$$f(\mathbf{A}) = \mathbf{V} f(\Sigma) \mathbf{V}^T , \tag{13.57}$$

with $f(\Sigma)$ given by

$$f(\Sigma) = \begin{pmatrix} f(\lambda_1) & 0 & \cdots & 0 \\ 0 & f(\lambda_2) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & & f(\lambda_N) \end{pmatrix} . \tag{13.58}$$

Problem f In order to revert to an explicit eigenvector expansion, show that (13.57) can be written as:

$$f(\mathbf{A}) = \sum_{n=1}^{N} f(\lambda_n) \hat{\mathbf{v}}^{(n)} \hat{\mathbf{v}}^{(n)T} . \tag{13.59}$$

With this expression (or the equivalent expression (13.57)) the evaluation of $f(\mathbf{A})$ is simple once the eigenvectors and eigenvalues of $\mathbf{A}$ are known, because in (13.59) the function f only acts on the eigenvalues, but not on the matrix. Since the function f normally acts on a scalar (such as the eigenvalues), the eigenvector decomposition has obviated the need for computing higher powers of the matrix $\mathbf{A}$. From a numerical point of view, however, computing functions of matrices can be a tricky issue. For example, Moler and van Loan [70] give nineteen dubious ways to compute the exponential of a matrix.

13.6 Normal modes of a vibrating system

An eigenvector decomposition is not only useful for computing the inverse of a matrix and other functions of a matrix, it also provides a way to analyze characteristics of dynamical systems. As an example, a simple model for the oscillations of a vibrating molecule is discussed here. This system is the prototype of a vibrating system that has different modes of vibration. The natural modes of vibration are

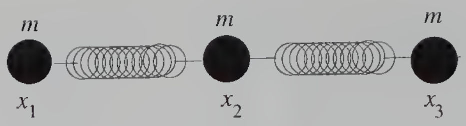

Fig. 13.6 Definition of variables for a simple vibrating system.

usually called the *normal modes* of that system. Consider the mechanical system shown in Figure 13.6. Three particles with mass m are coupled by two springs with spring constant k. It is assumed that the three masses are constrained to move along a line. The displacements of the masses from their equilibrium positions are denoted with x_1, x_2, and x_3. This mechanical model can considered to be a grossly oversimplified model of a tri-atomic molecule such as CO_2 or H_2O.

Each of the masses can experience an external force F_i, where the subscript i denotes the mass under consideration. The equations of motion for the three masses are given by:

$$\left.\begin{aligned} m\ddot{x}_1 &= k(x_2 - x_1) + F_1, \\ m\ddot{x}_2 &= -k(x_2 - x_1) + k(x_3 - x_2) + F_2, \\ m\ddot{x}_3 &= -k(x_3 - x_2) + F_3. \end{aligned}\right\} \tag{13.60}$$

For the moment we consider harmonic oscillations, that is we assume that both the driving forces F_i and the displacements x_i vary with time as $e^{-i\omega t}$. The displacements x_1, x_2, and x_3 can be used to form a vector $\mathbf{x}$, and similarly a vector $\mathbf{F}$ can be formed from the three forces F_1, F_2, and F_3 that act on the three masses.

Problem a Show that for a harmonic motion with frequency ω the equations of motion can be written in vector form as:

$$\left(\mathbf{A} - \frac{m\omega^2}{k}\mathbf{I}\right)\mathbf{x} = \frac{1}{k}\mathbf{F}, \tag{13.61}$$

with the matrix $\mathbf{A}$ given by

$$\mathbf{A} = \begin{pmatrix} 1 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 1 \end{pmatrix}. \tag{13.62}$$

The normal modes of the system are given by the patterns of oscillations of the system when there is no driving force. For this reason, we set the driving force $\mathbf{F}$ on the right-hand side of (13.61) momentarily to zero. Equation (13.61) then reduces to a homogeneous system of linear equations; such a system of equations can only have nonzero solutions when the determinant of the matrix vanishes. Since the matrix $\mathbf{A}$ has three eigenvalues, the system can only oscillate freely at three discrete eigenfrequencies. The system can only oscillate at other frequencies when it is driven by the force **F** at such a frequency.

Problem b Show that the eigenfrequencies ω_i of the vibrating system are given by

$$\omega_i = \sqrt{\frac{k\lambda_i}{m}}, \tag{13.63}$$

where λ_i are the eigenvalues of the matrix **A**.

Problem c Show that the eigenfrequencies of the system are given by:

$$\omega_1 = 0, \qquad \omega_2 = \sqrt{\frac{k}{m}}, \qquad \omega_3 = \sqrt{\frac{3k}{m}}. \tag{13.64}$$

Problem d The frequencies do not give the vibrations of each of the three particles respectively. Instead these frequencies give the eigenfrequencies of the three modes of oscillation of the system. The eigenvector that corresponds to each eigenvalue gives the displacement of each particle for that mode of oscillation. Show that these eigenvectors are given by:

$$\hat{\mathbf{v}}^{(1)} = \frac{1}{\sqrt{3}}\begin{pmatrix}1\\1\\1\end{pmatrix}, \qquad \hat{\mathbf{v}}^{(2)} = \frac{1}{\sqrt{2}}\begin{pmatrix}1\\0\\-1\end{pmatrix}, \qquad \hat{\mathbf{v}}^{(3)} = \frac{1}{\sqrt{6}}\begin{pmatrix}1\\-2\\1\end{pmatrix}. \tag{13.65}$$

Remember that the eigenvectors can be multiplied by an arbitrary constant, and this constant is chosen in such a way that each eigenvector has length 1.

Problem e Show that these eigenvectors satisfy the requirement (13.38).

Problem f Sketch the motion of the three masses for each normal mode. Explain physically why the third mode with frequency ω_3 has a higher eigenfrequency than the second mode ω_2.

Problem g Explain physically why the second mode has an eigenfrequency $\omega_2 = \sqrt{k/m}$ that is identical to the frequency of a single mass m that is suspended by a spring with spring constant k.

Problem h What type of motion does the first mode with eigenfrequency ω_1 describe? Explain physically why this frequency is independent of the spring constant k and the mass m.

Now that we know the normal modes of the system, we consider the case in which the system is driven by a force **F** that varies in time as $e^{-i\omega t}$. For simplicity it is assumed that the frequency ω of the driving force differs from the eigenfrequencies of the system: $\omega \neq \omega_i$. The eigenvectors $\hat{\mathbf{v}}^{(n)}$ defined in (13.65) form a complete orthonormal set, hence both the driving force **F** and the displacement **x** can be expanded in this set. Using (13.43) the driving force can be expanded as

$$\mathbf{F} = \sum_{n=1}^{3} \hat{\mathbf{v}}^{(n)} (\hat{\mathbf{v}}^{(n)} \cdot \mathbf{F}). \tag{13.66}$$

Problem i Write the displacement vector as a superposition of the normal mode displacements: $\mathbf{x} = \sum_{n=1}^{3} c_n \hat{\mathbf{v}}^{(n)}$, use expansion (13.66) for the driving force and insert these equations in the equation of motion (13.61) to solve for the unknown coefficients c_n. Eliminate the eigenvalues with (13.63) in favour of the eigenfrequencies ω_n, and show that the displacement is given by:

$$\mathbf{x} = \frac{1}{m} \sum_{n=1}^{3} \frac{\hat{\mathbf{v}}^{(n)} (\hat{\mathbf{v}}^{(n)} \cdot \mathbf{F})}{\left(\omega_n^2 - \omega^2\right)}. \tag{13.67}$$

This expression has a nice physical interpretation. Expression (13.67) states that the total response of the system can be written as a superposition of the different normal modes (the $\sum_{n=1}^{3} \hat{\mathbf{v}}^{(n)}$ terms). The effect that the force has on each normal mode is given by the inner product $(\hat{\mathbf{v}}^{(n)} \cdot \mathbf{F})$. This is nothing but the component of the force **F** along the eigenvector $\hat{\mathbf{v}}^{(n)}$, see (13.2). The term $1/\left(\omega_n^2 - \omega^2\right)$ gives the sensitivity of each mode to a driving force with frequency ω; this term can be called a sensitivity term. When the driving force is close to one of the eigenfrequencies of the nth mode, $1/\left(\omega_n^2 - \omega^2\right)$ is large. In that case the system is close to resonance and the resulting displacement is large. On the other hand, when the frequency of the driving force is far from the eigenfrequencies of the system, $1/\left(\omega_n^2 - \omega^2\right)$ is small and the system gives a small response. The total response can be seen as a combination of three basic operations: eigenvector expansion, projection, and multiplication with a response function. Note that the same operations were used in the explanation of the action of a matrix **A** given below equation (13.44).

13.7 Singular value decomposition

In Section 13.4 the decomposition of a square matrix in terms of eigenvectors was treated. In many practical applications, such as inverse problems, one encounters a system of equations that is not square:

$$\underbrace{\mathbf{A}}_{\substack{M\times N\\ \textit{matrix}}}\ \underbrace{\mathbf{x}}_{\substack{N\\ \textit{rows}}} = \underbrace{\mathbf{y}}_{\substack{M\\ \textit{rows}}} \tag{13.68}$$

Consider the example in which the vector **x** has N components and there are M equations. In that case the vector **y** has M components and the matrix **A** has M rows and N columns, that is it is an $M \times N$ matrix. A relation such as (13.37) which states that $\mathbf{A}\hat{\mathbf{v}}^{(n)} = \lambda_n \hat{\mathbf{v}}^{(n)}$ cannot possibly hold because when the matrix **A** acts on an N-vector it produces an M-vector, whereas in (13.37) the vector on the right-hand side has the same number of components as the vector on the left-hand side. It is clear that the theory of Section 13.4 cannot be applied when the matrix is not square. However, it is possible to generalize the theory of Section 13.4 when **A** is not square. For simplicity it is assumed that **A** is a real matrix.

In Section 13.4 a single set of orthonormal eigenvectors $\hat{\mathbf{v}}^{(n)}$ was used to analyze the problem. Since the vectors **x** and **y** in (13.68) have different dimensions, it is necessary to expand the vector **x** in a set of N orthogonal vectors $\hat{\mathbf{v}}^{(n)}$ that each have N components and to expand **y** in a different set of M orthogonal vectors $\hat{\mathbf{u}}^{(m)}$ that each have M components. Suppose we have chosen a set $\hat{\mathbf{v}}^{(n)}$, let us define vectors $\hat{\mathbf{u}}^{(n)}$ by the following relation:

$$\mathbf{A}\hat{\mathbf{v}}^{(n)} = \lambda_n \hat{\mathbf{u}}^{(n)}. \tag{13.69}$$

The constant λ_n should not be confused with an eigenvalue, this constant follows from the requirement that $\hat{\mathbf{v}}^{(n)}$ and $\hat{\mathbf{u}}^{(n)}$ are both vectors of unit length. At this point, the choice of $\hat{\mathbf{v}}^{(n)}$ is still open. The vectors $\hat{\mathbf{v}}^{(n)}$ will now be required to satisfy in addition to (13.69) the following condition:

$$\mathbf{A}^T \hat{\mathbf{u}}^{(n)} = \mu_n \hat{\mathbf{v}}^{(n)}, \tag{13.70}$$

where $\mathbf{A}^T$ is the transpose of **A**.

Problem a In order to find the vectors $\hat{\mathbf{v}}^{(n)}$ and $\hat{\mathbf{u}}^{(n)}$ that satisfy both (13.69) and (13.70), multiply (13.69) by $\mathbf{A}^T$ and use (13.70) to eliminate $\hat{\mathbf{u}}^{(n)}$. Do this to show that $\hat{\mathbf{v}}^{(n)}$ satisfies:

$$\left(\mathbf{A}^T\mathbf{A}\right)\hat{\mathbf{v}}^{(n)} = \lambda_n \mu_n \hat{\mathbf{v}}^{(n)}. \tag{13.71}$$

Use similar steps to show that $\hat{\mathbf{u}}^{(n)}$ satisfies

$$\left(\mathbf{A}\mathbf{A}^T\right)\hat{\mathbf{u}}^{(n)} = \lambda_n \mu_n \hat{\mathbf{u}}^{(n)}. \tag{13.72}$$

These equations state that the $\hat{\mathbf{v}}^{(n)}$ are the eigenvectors of $\mathbf{A}^T\mathbf{A}$ and that the $\hat{\mathbf{u}}^{(n)}$ are the eigenvectors of $\mathbf{A}\mathbf{A}^T$.

Problem b Show that both $\mathbf{A}^T\mathbf{A}$ and $\mathbf{A}\mathbf{A}^T$ are real symmetric matrices and that this implies that the basis vectors $\hat{\mathbf{v}}^{(n)}$ $(n = 1, \ldots, N)$ and $\hat{\mathbf{u}}^{(m)}$ $(m = 1, \ldots, M)$ are both orthonormal:

$$\left(\hat{\mathbf{v}}^{(n)} \cdot \hat{\mathbf{v}}^{(m)}\right) = \left(\hat{\mathbf{u}}^{(n)} \cdot \hat{\mathbf{u}}^{(m)}\right) = \delta_{nm}. \tag{13.73}$$

Although (13.71) and (13.72) can be used to find the basis vectors $\hat{\mathbf{v}}^{(n)}$ and $\hat{\mathbf{u}}^{(n)}$, these expressions cannot be used to find the constants λ_n and μ_n, because they state that the *product* $\lambda_n\mu_n$ is equal to the eigenvalues of $\mathbf{A}^T\mathbf{A}$ and $\mathbf{A}\mathbf{A}^T$. This implies that only the product of λ_n and μ_n is defined.

Problem c In order to find the relation between λ_n and μ_n, take the inner product of (13.69) with $\hat{\mathbf{u}}^{(n)}$ and use the orthogonality relation (13.73) to show that:

$$\lambda_n = \left(\hat{\mathbf{u}}^{(n)} \cdot \mathbf{A}\hat{\mathbf{v}}^{(n)}\right). \tag{13.74}$$

Problem d Show that for arbitrary vectors $\mathbf{p}$ and $\mathbf{q}$

$$(\mathbf{p} \cdot \mathbf{A}\mathbf{q}) = \left(\mathbf{A}^T\mathbf{p} \cdot \mathbf{q}\right). \tag{13.75}$$

Problem e Apply this relation to (13.74) and use (13.70) to show that

$$\lambda_n = \mu_n. \tag{13.76}$$

This is all the information we need to find both λ_n and μ_n. Since these quantities are equal, and since by virtue of (13.71) these eigenvectors are equal to the eigenvectors of $\mathbf{A}^T\mathbf{A}$, it follows that both λ_n and μ_n are given by the square-root of the eigenvalues of $\mathbf{A}^T\mathbf{A}$. Note that it follows from (13.72) that the product $\lambda_n\mu_n$ also equals the eigenvalues of $\mathbf{A}\mathbf{A}^T$. This can only be the case when $\mathbf{A}^T\mathbf{A}$ and $\mathbf{A}\mathbf{A}^T$ have the same eigenvalues. Before we proceed, let us show that this is indeed true. Let the eigenvalues of $\mathbf{A}^T\mathbf{A}$ be denoted by Λ_n and the eigenvalues of $\mathbf{A}\mathbf{A}^T$ by Υ_n, that is

$$\mathbf{A}^T\mathbf{A}\hat{\mathbf{v}}^{(n)} = \Lambda_n\hat{\mathbf{v}}^{(n)} \tag{13.77}$$

and

$$\mathbf{A}\mathbf{A}^T\hat{\mathbf{u}}^{(n)} = \Upsilon_n\hat{\mathbf{u}}^{(n)}. \tag{13.78}$$

Problem f Take the inner product of (13.77) with $\hat{\mathbf{v}}^{(n)}$ to show that $\Lambda_n = (\hat{\mathbf{v}}^{(n)} \cdot \mathbf{A}^T\mathbf{A}\hat{\mathbf{v}}^{(n)})$, use the properties (13.75) and $\mathbf{A}^{TT} = \mathbf{A}$ and (13.69) to show that $\lambda_n^2 = \Lambda_n$. Use similar steps to show that $\mu_n^2 = \Upsilon_n$. With (13.76) this implies that $\mathbf{AA}^T$ and $\mathbf{A}^T\mathbf{A}$ have the same eigenvalues.

The proof that $\mathbf{AA}^T$ and $\mathbf{A}^T\mathbf{A}$ have the same eigenvalues was not only given as a check of the consistency of the theory, the fact that $\mathbf{AA}^T$ and $\mathbf{A}^T\mathbf{A}$ have the same eigenvalues has important implications. Since $\mathbf{AA}^T$ is an $M \times M$ matrix, it has M eigenvalues, and since $\mathbf{A}^T\mathbf{A}$ is an $N \times N$ matrix it has N eigenvalues. The only way for these matrices to have the same eigenvalues, but to have a different number of eigenvalues is for the number of nonzero eigenvalues to be given by the minimum of N and M. In practice, some of the eigenvalues of $\mathbf{AA}^T$ may be zero, hence the number of nonzero eigenvalues of $\mathbf{AA}^T$ can be less than M. By the same token, the number of nonzero eigenvalues of $\mathbf{A}^T\mathbf{A}$ can be less than N. The number of nonzero eigenvalues will be denoted by P. It is not known a priori how many nonzero eigenvalues there are, but it follows from the arguments above that P is smaller than or equal to M and N. This implies that

$$P \le \min(N, M), \tag{13.79}$$

where $\min(N, M)$ denotes the minimum of N and M. Therefore, whenever a summation over eigenvalues occurs, we need to take only P eigenvalues into account. Since the ordering of the eigenvalues is arbitrary, it is assumed in the following that the eigenvectors are ordered in decreasing size: $\lambda_1 \ge \lambda_2 \ge \cdots \ge \lambda_P$. In this ordering the eigenvalues for $n > P$ are equal to zero so that the summation over eigenvalues runs from 1 to P.

Problem g The matrices $\mathbf{AA}^T$ and $\mathbf{A}^T\mathbf{A}$ have the same eigenvalues. When you need the eigenvalues and eigenvectors, from the point of view of computational efficiency would it be more efficient to compute the eigenvalues and eigenvectors of $\mathbf{A}^T\mathbf{A}$ or of $\mathbf{AA}^T$? Consider the situations $M > N$ and $M < N$ separately.

Let us now return to the task of making an eigenvalue decomposition of the matrix $\mathbf{A}$. The vectors $\hat{\mathbf{v}}^{(n)}$ form a basis in N-dimensional space. Since the vector $\mathbf{x}$ is N-dimensional, every vector $\mathbf{x}$ can be decomposed according to (13.43): $\mathbf{x} = \sum_{n=1}^{N} \hat{\mathbf{v}}^{(n)}(\hat{\mathbf{v}}^{(n)} \cdot \mathbf{x})$.

Problem h Let the matrix **A** act on this expression and use (13.69) to show that:

$$\mathbf{Ax} = \sum_{n=1}^{P} \lambda_n \hat{\mathbf{u}}^{(n)} \left(\hat{\mathbf{v}}^{(n)} \cdot \mathbf{x} \right). \tag{13.80}$$

Problem i This expression must hold for any vector **x**. Use this property to deduce that:

$$\mathbf{A} = \sum_{n=1}^{P} \lambda_n \hat{\mathbf{u}}^{(n)} \hat{\mathbf{v}}^{(n)T}. \tag{13.81}$$

Problem j The eigenvectors $\hat{\mathbf{v}}^{(n)}$ can be arranged in an $N \times N$ matrix **V**, defined in (13.39). Similarly the eigenvectors $\hat{\mathbf{u}}^{(n)}$ can be used to form the columns of an $M \times M$ matrix **U**:

$$\mathbf{U} = \begin{pmatrix} \vdots & \vdots & & \vdots \\ \hat{\mathbf{u}}^{(1)} & \hat{\mathbf{u}}^{(2)} & \cdots & \hat{\mathbf{u}}^{(M)} \\ \vdots & \vdots & & \vdots \end{pmatrix}. \tag{13.82}$$

Show that **A** can also be written as:

$$\mathbf{A} = \mathbf{U}\Sigma\mathbf{V}^T, \tag{13.83}$$

with the diagonal matrix Σ defined in (13.47).

This decomposition of **A** in terms of eigenvectors is called the *singular value decomposition* of the matrix. This is frequently abbreviated to SVD.

Problem k You may have noticed the similarity between (13.81) and (13.45) for a square matrix and (13.83) and (13.46). Show that for the special case $M = N$ the theory of this section is identical to the eigenvalue decomposition for a square matrix presented in Section 13.4. Hint: what are the vectors $\hat{\mathbf{u}}^{(n)}$ when $M = N$?

Let us now solve the original system of linear equations (13.68) for the unknown vector **x**. In order to do this, expand the vector **y** in the vectors $\hat{\mathbf{u}}^{(n)}$ that span the M-dimensional space: $\mathbf{y} = \sum_{m=1}^{M} \hat{\mathbf{u}}^{(m)} (\hat{\mathbf{u}}^{(m)} \cdot \mathbf{y})$, and expand the vector **x** in the vectors $\hat{\mathbf{v}}^{(n)}$ that span the N-dimensional space:

$$\mathbf{x} = \sum_{n=1}^{N} c_n \hat{\mathbf{v}}^{(n)}. \tag{13.84}$$

Problem l At this point the coefficients c_n are unknown. Insert the expansions for $\mathbf{y}$ and $\mathbf{x}$ and the expansion (13.81) for the matrix $\mathbf{A}$ in the linear system (13.68) and use the orthogonality properties of the eigenvectors to show that $c_n = (\hat{\mathbf{u}}^{(n)} \cdot \mathbf{y})/\lambda_n$, so that

$$\mathbf{x} = \sum_{n=1}^{P} \frac{1}{\lambda_n} \left(\hat{\mathbf{u}}^{(n)} \cdot \mathbf{y}\right) \hat{\mathbf{v}}^{(n)}. \tag{13.85}$$

Note that although in the original expansion (13.84) a summation is carried out over all N basis vectors, in solution (13.85) a summation is carried out over the first P basis vectors only. The reason for this is that the remaining eigenvectors have eigenvalues that are equal to zero so that they can be left out of the expansion (13.81) of the matrix $\mathbf{A}$. Indeed, these eigenvalues would give rise to problems because if they were retained they would lead to infinite contributions $1/\lambda \to \infty$ in solution (13.85). In practice, some eigenvalues may be nonzero, but close to zero, so that the term $1/\lambda$ gives rise to numerical instabilities. Therefore, one also often leaves out nonzero but small eigenvalues in summation (13.85).

This may appear to be an objective procedure for defining solutions for linear problems that are undetermined or for problems that are otherwise ill-conditioned, but there is a price one pays for leaving out basis vectors in the construction of the solution. The vector $\mathbf{x}$ is N-dimensional, hence one needs N basis vectors to construct an arbitrary vector $\mathbf{x}$, see (13.84). The solution vector given in (13.85) is built by superposing only P basis vectors. This implies that the solution vector is constrained to be within the P-dimensional subspace spanned by the first P eigenvectors. Therefore, it is not clear that the solution vector in (13.85) is identical to the true vector $\mathbf{x}$. However, the point of using the singular value decomposition is that the solution is only constrained by the linear system of (13.68) within the subspace spanned by the first P basis vectors $\hat{\mathbf{v}}^{(n)}$. Solution (13.85) ensures that only the components of $\mathbf{x}$ within that subspace are affected by the right-hand side vector $\mathbf{y}$. This technique is useful in the analysis of linear inverse problems [79].

13.8 Householder transformation

Linear systems of equations can be solved in a systematic way by sweeping columns of the matrix that defines the system of equations. As an example consider the system

$$\left.\begin{aligned} x + y + 3z &= 5, \\ -x + 2z &= 1, \\ 2x + y + 2z &= 5. \end{aligned}\right\} \tag{13.86}$$

This system of equations will be written here also as:

$$\left(\begin{array}{ccc|c} 1 & 1 & 3 & 5 \\ -1 & 0 & 2 & 1 \\ 2 & 1 & 2 & 5 \end{array}\right), \tag{13.87}$$

which is nothing but a compressed notation of (13.86). The matrix shown in (13.87) is called the *augmented matrix* because the matrix defining the left-hand side of (13.86) is augmented with the right-hand side of (13.86). The linear equations can be solved by adding the first row to the second row and subtracting the first row twice from the third row, the resulting system of equations is then represented by the following augmented matrix:

$$\left(\begin{array}{ccc|c} 1 & 1 & 3 & 5 \\ 0 & 1 & 5 & 6 \\ 0 & -1 & -4 & -5 \end{array}\right). \tag{13.88}$$

Note that in the first column all elements below the first elements are equal to zero. By adding the second row to the third row we can also make all elements below the second element in the second column equal to zero:

$$\left(\begin{array}{ccc|c} 1 & 1 & 3 & 5 \\ 0 & 1 & 5 & 6 \\ 0 & 0 & 1 & 1 \end{array}\right). \tag{13.89}$$

The system is now in *upper-triangular form*; this is a different way of saying that all matrix elements below the diagonal vanish. This is convenient because the system can now be solved by *backsubstitution*. To see how this works note that the augmented matrix (13.89) is a shorthand notation for the following system of equations:

$$\left.\begin{aligned} x + y + 3z &= 5, \\ y + 5z &= 6, \\ z &= 1\,. \end{aligned}\right\} \tag{13.90}$$

The value of z follows from the last equation, given this value of z the value of y follows from the middle equations, and given y and z the value of x follows from the top equation.

Problem a Show that the solution of the linear equations is given by $x = y = z = 1$.

For small systems of linear equations this process for solving linear equations can be carried out by hand. For large systems of equations this process must be carried out on a computer. This is only possible when one has a systematic and efficient way of carrying out this sweeping process. Suppose we have an $N \times N$ matrix $\mathbf{A}$:

$$\mathbf{A} = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1N} \\ a_{21} & a_{22} & \cdots & a_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ a_{N1} & a_{N2} & \cdots & a_{NN} \end{pmatrix}. \tag{13.91}$$

We want to find an operator $\mathbf{Q}$ such that when $\mathbf{A}$ is multiplied by $\mathbf{Q}$ all elements in the first column are zero except the element on the diagonal, that is we want to find $\mathbf{Q}$ such that:

$$\mathbf{QA} = \begin{pmatrix} a'_{11} & a'_{12} & \cdots & a'_{1N} \\ 0 & a'_{22} & \cdots & a'_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & a'_{N2} & \cdots & a'_{NN} \end{pmatrix}. \tag{13.92}$$

This problem can be formulated slightly differently, suppose we denote the first columns of $\mathbf{A}$ by the vector $\mathbf{u}$:

$$\mathbf{u} \equiv \begin{pmatrix} a_{11} \\ a_{21} \\ \vdots \\ a_{N1} \end{pmatrix}. \tag{13.93}$$

The operator $\mathbf{Q}$ that we want to find maps this vector to a new vector which only has a nonzero component in the first element:

$$\mathbf{Qu} = \begin{pmatrix} a'_{11} \\ 0 \\ \vdots \\ 0 \end{pmatrix} = a'_{11}\hat{\mathbf{e}}_1, \tag{13.94}$$

where $\hat{\mathbf{e}}_1$ is the unit vector in the x_1-direction:

$$\hat{\mathbf{e}}_1 \equiv \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}. \tag{13.95}$$

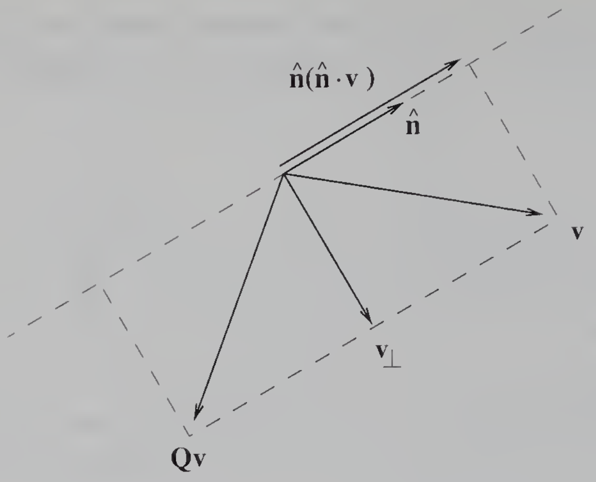

Fig. 13.7 Geometrical interpretation of the Householder transformation.

The desired operator **Q** can be found with a so-called Householder transformation. For a given unit vector $\hat{\mathbf{n}}$ the Householder transformation is defined by:

$$\mathbf{Q} \equiv \mathbf{I} - 2\hat{\mathbf{n}}\hat{\mathbf{n}}^T . \tag{13.96}$$

Problem b Show that the Householder transformation can be written as $\mathbf{Q} = \mathbf{I} - 2\mathbf{P}$, where **P** is the operator for projection along $\hat{\mathbf{n}}$.

Problem c It follows from (13.3) that any vector **v** can be decomposed into a component along $\hat{\mathbf{n}}$ and a perpendicular component: $\mathbf{v} = \hat{\mathbf{n}}\,(\hat{\mathbf{n}} \cdot \mathbf{v}) + \mathbf{v}_\perp$. Show that after the Householder transformation the vector is given by:

$$\mathbf{Qv} = -\hat{\mathbf{n}}\,(\hat{\mathbf{n}} \cdot \mathbf{v}) + \mathbf{v}_\perp . \tag{13.97}$$

Problem d Convince yourself that the Householder transformation of **v** is correctly shown in Figure 13.7.

Problem e Use (13.97) to show that **Q** does not change the length of a vector. Use this result to show that a'_{11} in (13.94) is given by $a'_{11} = |\mathbf{u}|$.

With (13.94) this means that the Householder transformation should satisfy

$$\mathbf{Qu} = |\mathbf{u}|\,\hat{\mathbf{e}}_1 . \tag{13.98}$$

Our goal is now to find a unit vector $\hat{\mathbf{n}}$ such that this expression is satisfied.

Problem f Use (13.96) to show that if $\mathbf{Q}$ satisfies the requirement (13.98) $\hat{\mathbf{n}}$ must satisfy the following equation:

$$2\hat{\mathbf{n}}\,(\hat{\mathbf{n}}\cdot\hat{\mathbf{u}}) = \hat{\mathbf{u}} - \hat{\mathbf{e}}_1; \tag{13.99}$$

in this expression $\hat{\mathbf{u}}$ is the unit vector in the direction $\mathbf{u}$.

Problem g Equation (13.99) implies that $\hat{\mathbf{n}}$ is directed in the direction of the vector $\hat{\mathbf{u}} - \hat{\mathbf{e}}_1$, therefore $\hat{\mathbf{n}}$ can be written as $\hat{\mathbf{n}} = C\,(\hat{\mathbf{u}} - \hat{\mathbf{e}}_1)$, with C an undetermined constant. Show that (13.98) implies that $C = 1/\sqrt{2\,[1-(\hat{\mathbf{u}}\cdot\hat{\mathbf{e}}_1)]}$. Also show that this value of C indeed leads to a vector $\hat{\mathbf{n}}$ that is of unit length.

This value of C implies that the unit vector $\hat{\mathbf{n}}$ to be used in the Householder transformation (13.96) is given by

$$\hat{\mathbf{n}} = \frac{\hat{\mathbf{u}} - \hat{\mathbf{e}}_1}{\sqrt{2\,[1-(\hat{\mathbf{u}}\cdot\hat{\mathbf{e}}_1)]}}. \tag{13.100}$$

To see how the Householder transformation can be used to render the matrix elements below the diagonal equal to zero apply the transformation $\mathbf{Q}$ to the linear equation $\mathbf{Ax} = \mathbf{y}$.

Problem h Show that this leads to a new system of equations given by

$$\begin{pmatrix} |\mathbf{u}| & a'_{12} & \cdots & a'_{1N} \\ 0 & a'_{22} & \cdots & a'_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & a'_{N2} & \cdots & a'_{NN} \end{pmatrix} \mathbf{x} = \mathbf{Qy}. \tag{13.101}$$

A second Householder transformation can now be applied to render all elements in the second column below the diagonal element a'_{22} equal to zero. In this way, all the columns of $\mathbf{A}$ can be successively swiped. Note that in order to apply the Householder transformation one only needs to compute (13.100) and (13.96) to carry out a matrix multiplication. These operations can be carried out efficiently on computers.

14

Dirac delta function

14.1 Introduction of the delta function

In linear algebra, the identity matrix **I** plays a central role. This operator maps any vector **v** onto itself:

$$\mathbf{I}\mathbf{v} = \mathbf{v}. \tag{14.1}$$

This expression can also be written in component form as

$$\sum_j I_{ij} v_j = v_i. \tag{14.2}$$

The identity matrix has diagonal elements that are equal to unity and its off-diagonal elements are equal to zero. This means that the elements of the identity matrix are equal to the Kronecker delta $I_{ij} = \delta_{ij}$, which is defined as:

$$\delta_{ij} \equiv \begin{cases} 1 & \text{for} \quad i = j \\ 0 & \text{for} \quad i \neq j \end{cases}. \tag{14.3}$$

Expression (14.2) shows that when the identity matrix I_{ij} acts on all the components v_j of a vector, it selects the component v_i. The question we address in this chapter is: how can this idea be generalized to continuous functions instead of vectors? In other words, can we find a function $I(x_0, x)$ such that when it is integrated with a function $f(x)$ it selects that function as the location x_0:

$$\int I(x_0, x) f(x) dx = f(x_0)? \tag{14.4}$$

Note the resemblance between this expression and (14.2) for a vector. The vector **v** is replaced by the function f, the summation over j is changed into the integration over x, and the index i in (14.2) corresponds to the value x_0 in (14.4).

As a first guess for the operator $I(x_0, x)$ let us consider the following generalization of the definition (14.3) of the Kronecker delta to continuous functions:

$$I(x_0, x) \overset{?}{\equiv} \begin{cases} 1 & \text{for} \quad x_0 = x \\ 0 & \text{for} \quad x_0 \neq x \end{cases} . \tag{14.5}$$

This is not a very good guess; when $I(x_0, x)$ is viewed as a function of x, $I(x_0, x)$ is almost everywhere equal to zero except when $x = x_0$. When we integrate $I(x_0, x)$ over x the integrand is zero except at the point $x = x_0$, but this point gives a vanishing contribution to the integral because the "width" of this point is equal to zero. (Mathematically one would say that a point has zero measure.) This means that the integral of $I(x_0, x)$ can only give a finite result when $I(x_0, x)$ is infinite at $x = x_0$.

As an improved guess let us therefore try the definition

$$I(x_0, x) \overset{?}{\equiv} \begin{cases} \infty & \text{for} \quad x_0 = x \\ 0 & \text{for} \quad x_0 \neq x \end{cases} . \tag{14.6}$$

This is not a very precise definition because it is not clear what we mean by "∞". However, we can learn something from this naive guess because it shows that $I(x_0, x)$ is not a well-behaved function. It is discontinuous at $x = x_0$ and its value is infinite at that point. It is clear from this that whatever definition of $I(x_0, x)$ we use, it will not lead to a well-behaved function.

A useful definition of $I(x_0, x)$ can be obtained from the boxcar function $B_a(x)$ which is defined as

$$B_a(x) \equiv \begin{cases} \dfrac{1}{2a} & \text{for} \quad |x| \leq a \\ 0 & \text{for} \quad |x| > a \end{cases} . \tag{14.7}$$

This function is shown for three values of the parameter a in Figure 14.1. (A boxcar is the rectangular railroad car that is used for carrying freight, and the function $B_a(x)$

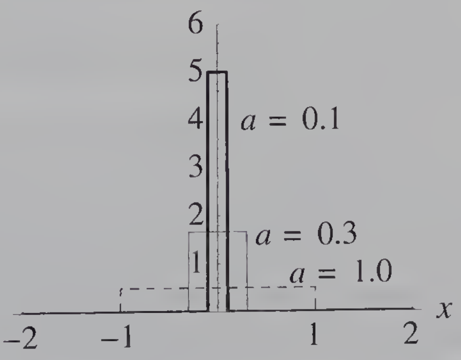

Fig. 14.1 The boxcar function $B_a(x)$ for $a = 1.0$ (dashed line), $a = 0.3$ (thin solid line), and $a = 0.1$ (thick solid line).

is called the boxcar function because it has the same rectangular shape as the boxcar used by the railways.)

Problem a Show that

$$\int_{-\infty}^{\infty} B_a(x)dx = 1. \tag{14.8}$$

Let us now center the boxcar at a location x_0, multiply it by $f(x)$ and integrate over x.

Problem b Use the definition of the boxcar function to derive that

$$\int_{-\infty}^{\infty} B_a(x - x_0)f(x)dx = \frac{1}{2a}\int_{x_0-a}^{x_0+a} f(x)\,dx. \tag{14.9}$$

Pay attention to the limits of integration on the right-hand side.

The integral on the right-hand side is nothing but the average of the function $f(x)$ over the interval $(x_0 - a, x_0 + a)$ because the pre-factor $1/2a$ corrects for the width of that interval. This is a useful expression; as the parameter a goes to zero the integral (14.9) gives the function at location x_0 because the limit $a \downarrow 0$ gives the mean value of $f(x)$ over the interval $(x_0 - 0, x_0 + 0)$. This means that

$$\lim_{a\downarrow 0}\int_{-\infty}^{\infty} B_a(x - x_0)f(x)\,dx = f(x_0). \tag{14.10}$$

This means that this limit has the desired properties of the identity operator $I(x_0, x)$ for continuous functions. It is customary to denote this operator as $\delta(x - x_0)$ and to call it the *Dirac delta function*. (Usually one refers to this function simply as the "delta function.") This means that the delta function satisfies the following property:

$$\int_{-\infty}^{\infty} \delta(x - x_0)f(x)\,dx = f(x_0). \tag{14.11}$$

A comparison of this expression with (14.10) suggests that

$$\delta(x - x_0)\text{“}=\text{”}\lim_{a\downarrow 0} B_a(x - x_0). \tag{14.12}$$

Consider Figure 14.1 again. It can be seen from that figure and definition (14.7) that as a goes to zero, the value of the boxcar function becomes infinite and that the width of the boxcar goes to zero. In this sense the limit on the right-hand side does not exist, and for this reason the = sign is placed between quotes. The word "delta function" is really a misnomer because this "function" is not a function at all. Its value is not defined and it is only nonzero in a region with measure zero. However, within the integral (14.11) the action of the delta function is well defined, it is an operator that selects the function value $f(x_0)$ at position x_0. The delta function is an example of a *distribution*. This is a mathematical object that is only defined within an integral. Note that the limit $a \downarrow 0$ of the integral (14.10) *is* well defined. This means that the properties of the delta function can only be meaningfully stated when the delta function is used in an integral.

Let us first give a formal proof that definition (14.12) indeed leads to a delta function with the desired property (14.11). It is clear from (14.9) that we only need to consider the function $f(x)$ in the interval $(x_0 - a, x_0 + a)$ for small values of a. Therefore it is useful to represent the function $f(x)$ by a Taylor series around the point x_0.

Problem c Use (3.17) to show that $f(x)$ can be represented by the following Taylor series around the point x_0:

$$\begin{aligned} f(x) &= \sum_{n=0}^{\infty} \frac{1}{n!} \frac{d^n f}{dx^n}(x_0)(x - x_0)^n \\ &= f(x_0) + \frac{df}{dx}(x_0)(x - x_0) + \frac{1}{2} \frac{d^2 f}{dx^2}(x_0)(x - x_0)^2 + \cdots . \end{aligned} \qquad (14.13)$$

Problem d When this is inserted in (14.9) each term gives a contribution proportional to $\int_{x_0-a}^{x_0+a} (x - x_0)^n dx$. Show that for odd values of n this integral is equal to zero and that for even powers of n it is given by

$$\int_{x_0-a}^{x_0+a} (x - x_0)^n dx = \frac{2}{n+1} a^{n+1}. \qquad (14.14)$$

Problem e Use these results to show that

$$\begin{aligned} \int_{-\infty}^{\infty} B_a(x - x_0) f(x)\, dx &= \sum_{n\ even} \frac{1}{(n+1)!} \frac{d^n f}{dx^n}(x_0)\, a^n \\ &= f(x_0) + \frac{1}{6} \frac{d^2 f}{dx^2}(x_0)\, a^2 \\ &\quad + \frac{1}{120} \frac{d^4 f}{dx^4}(x_0)\, a^4 + \cdots . \end{aligned} \qquad (14.15)$$

In the limit $a \downarrow 0$ all the terms in this series vanish except the first term. This means that we have proven that (14.10) is indeed satisfied.

It should be noted that it is not necessary to define the delta function as the limit $a \downarrow 0$ of the boxcar function. One can also define the delta function using a

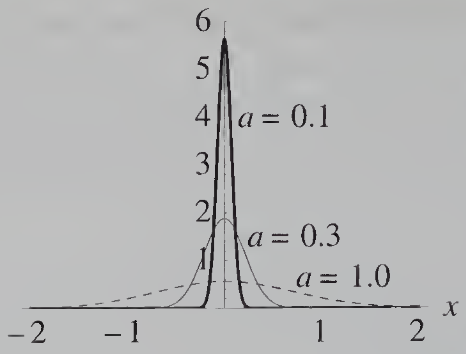

Fig. 14.2 The Gaussian function $g_a(x)$ for $a = 1.0$ (dashed line), $a = 0.3$ (thin solid line), and $a = 0.1$ (thick solid line).

Gaussian function with width a:

$$g_a(x) = \frac{1}{a\sqrt{\pi}} e^{-x^2/a^2}. \tag{14.16}$$

This function is shown for various values of a in Figure 14.2. A comparison of this figure with Figure 14.1 shows that for small values of a these functions have similar properties. One can indeed formally define the delta function using the Gaussian function (14.16) instead of the boxcar function. When the analysis of problem d is applied to this function, the first term is again given by $f(x_0)$. The higher-order terms have different coefficients because they follow from the integral $\int_{-\infty}^{\infty}(x - x_0)^n e^{-x^2/a^2} dx$ rather than the integral $\int_{x_0-a}^{x_0+a}(x - x_0)^n dx$. However, in the limit $a \downarrow 0$ these higher-order terms do not contribute. This example shows that the delta function can be defined as the limit of either boxcar functions or Gaussian functions. In fact the delta function can be defined as the limit of other types of functions as well.

14.2 Properties of the delta function

In the previous section the delta function was formally introduced. In this section we derive some properties of the delta function that are useful in a variety of applications.

Problem a Apply the identity (14.11) to the function $f(x) = 1$ to derive that

$$\int_{-\infty}^{\infty} \delta(x - x_0)\, dx = 1. \tag{14.17}$$

This expression states that the "surface area" under the delta function is equal to zero. As the width of this function goes to zero, the value of the function must become infinite to ensure that (14.17) is satisfied. Note that according to (14.8) the boxcar function $B_a(x)$ that we used to define the delta function indeed satisfies this property.

Problem b Show that the integral of the Gaussian function defined in (14.16) is also equal to unity:

$$\int_{-\infty}^{\infty} g_a(x)\,dx = 1. \tag{14.18}$$

In this derivation you can use that $\int_{-\infty}^{\infty} e^{-y^2} dy = \sqrt{\pi}$. This means that when the delta function is defined as the limit $a \downarrow 0$ of the Gaussian function $g_a(x)$ property (14.17) is indeed satisfied.

For the next property we consider the delta function $\delta\,(c(x - x_0))$, where c is a constant. Let us first consider the case in which c is positive.

Problem c Make the substitution $y = cx$ to derive the following identity:

$$\int_{-\infty}^{\infty} \delta\,(c\,(x - x_0))\,f(x)\,dx = \int_{-\infty}^{\infty} \delta\,(y - cx_0)\,f(y/c)\frac{1}{c}dy. \tag{14.19}$$

Problem d Carry out the y-integration using property (14.11) of the delta function to show that

$$\int_{-\infty}^{\infty} \delta\,(c\,(x - x_0))\,f(x)\,dx = \frac{1}{c}f(x_0) \qquad \text{(positive } c\text{).} \tag{14.20}$$

Problem e Carry out the same analysis for negative values of the constant c, and show that (14.19) in that case is given by

$$\int_{-\infty}^{\infty} \delta\,(c\,(x - x_0))\,f(x)\,dx = \int_{+\infty}^{-\infty} \delta\,(y - cx_0)\,f(y/c)\frac{1}{c}dy. \tag{14.21}$$

Explain why the integration runs from $+\infty$ to $-\infty$.

Problem f When the integration limits in the last integral are reversed, the integral obtains an additional minus sign. Carry out the y-integration and derive that

$$\int_{-\infty}^{\infty} \delta\,(c\,(x - x_0))\,f(x)\,dx = -\frac{1}{c}f(x_0) \qquad \text{(negative } c\text{).} \tag{14.22}$$

For negative values of c, one can use that $-c = |c|$. For positive values of c, obviously $c = |c|$. This means that (14.20) and (14.21) can be combined in the single property

$$\int_{-\infty}^{\infty} \delta\left(c\left(x - x_0\right)\right) f(x)\, dx = \frac{1}{|c|} f(x_0). \tag{14.23}$$

Following (14.11) we also know that the right-hand side of this expression can also be written as $(1/|c|)\int_{-\infty}^{\infty} \delta\left(x - x_0\right) f(x) dx$. A comparison with the left-hand side of (14.23) implies that

$$\delta\left(c\left(x - x_0\right)\right) = \frac{1}{|c|}\delta\left(x - x_0\right). \tag{14.24}$$

14.3 Delta function of a function

In some applications one arrives at an integral of the delta function in which the argument of the delta function is a function $g(x)$ rather than the integration variable x. This means that one needs to evaluate the integral $\int \delta\left(g(x)\right) f(x) dx$. We will encounter such an integral in (19.51). The delta function $\delta(x - x_0)$ only gives a nonzero contribution when x is close to x_0. Therefore, the delta function $\delta\left(g(x)\right)$ only needs to be evaluated near the value of x where $g(x) = 0$. Let us denote this value by x_0, so that

$$g(x_0) = 0. \tag{14.25}$$

This point x_0 is shown in Figure 14.3. Near the point x_0, the function $g(x)$ can be represented by a Taylor series:

$$g(x) = g(x_0) + \frac{dg}{dx}(x_0)(x - x_0) + \frac{1}{2}\frac{d^2 g}{dx^2}(x_0)(x - x_0)^2 + \cdots. \tag{14.26}$$

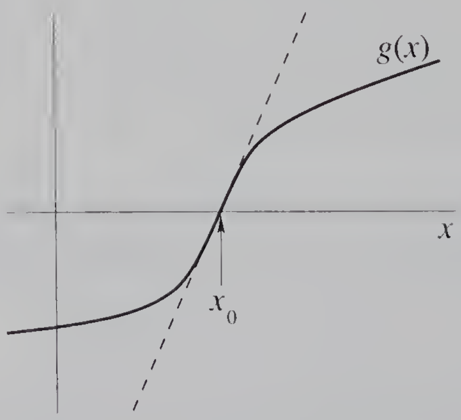

Fig. 14.3 The function $g(x)$ (thick solid line) with a zero-crossing at x_0 and the tangent line (dashed line) at that point.

Because x_0 gives the zero-crossing of $g(x)$, see (14.25), the first term is equal to zero. Ignoring the second order term and all higher order terms then gives the following first order approximation for the function $g(x)$ near its zero-crossing:

$$g(x) = \frac{dg}{dx}(x_0)(x - x_0). \tag{14.27}$$

Problem a Show that the right-hand side of (14.27) describes the straight line that is tangent to the curve $g(x)$ at the zero crossing x_0. This tangent line is shown as the dashed line in Figure 14.3.

Problem b Using this relation one finds that

$$\int \delta\left(g(x)\right) f(x)\, dx = \int \delta\left(\frac{dg}{dx}(x_0)(x - x_0)\right) f(x)\, dx. \tag{14.28}$$

The derivative dg/dx at the point x_0 can be considered to be a constant. Use the results of the previous section to derive that

$$\int \delta\left(g(x)\right) f(x)\, dx = \int \frac{1}{\left|\frac{dg}{dx}(x_0)\right|}\delta(x - x_0) f(x)\, dx, \tag{14.29}$$

and show that this gives

$$\int \delta\left(g(x)\right) f(x)\, dx = \frac{1}{\left|\frac{dg}{dx}(x_0)\right|} f(x_0). \tag{14.30}$$

Problem c Insert (14.11) into the right-hand side of (14.30) to derive the following property of the delta function

$$\delta\left(g(x)\right) = \frac{1}{\left|\frac{dg}{dx}(x_0)\right|}\delta(x - x_0), \tag{14.31}$$

where it must be kept in mind that x_0 denotes the zero-crossing of $g(x)$.

When the function $g(x)$ has more than one zero-crossing, the analysis of this section can be applied to each of these zero-crossings. The contributions of all the different zero-crossings must be added because all the points where $g(x) = 0$ give a contribution to the integral. This means that when the zero-crossings are denoted by x_i (so that $g(x_i) = 0$):

$$\delta(g(x)) = \sum_i \frac{1}{\left|\frac{dg}{dx}(x_i)\right|}\delta(x - x_i). \tag{14.32}$$

14.4 Delta function in more dimensions

The delta function has been defined for functions of a single variable. Its definition can be extended to functions of more variables. As an example we consider here the delta function in three dimensions. The delta function is then defined as the product of the delta functions for each of the coordinates:

$$\delta(\mathbf{r} - \mathbf{r}_0) \equiv \delta(x - x_0)\,\delta(y - y_0)\,\delta(z - z_0)\,. \tag{14.33}$$

This definition of the delta function can be used in the integral $\int \delta(\mathbf{r} - \mathbf{r}_0) f(\mathbf{r})\, dV$.

Problem a Write the volume integral as $dV = dxdydz$, insert (14.33) and carry out the integration over x to show that

$$\int \delta(\mathbf{r} - \mathbf{r}_0) f(\mathbf{r})\, dV = \int \delta(y - y_0)\,\delta(z - z_0)\, f(x_0, y, z)\, dydz, \tag{14.34}$$

paying attention to the arguments of the function f.

Problem b Carry out the y-integration and then the z-integration to derive that

$$\int \delta(\mathbf{r} - \mathbf{r}_0) f(\mathbf{r})\, dV = f(x_0, y_0, z_0). \tag{14.35}$$

The right-hand side of (14.35) is the function f at location $\mathbf{r}_0$. This means that the multi-dimensional delta function defined in (14.33) satisfies the following property

$$\int \delta(\mathbf{r} - \mathbf{r}_0) f(\mathbf{r})\, dV = f(\mathbf{r}_0). \tag{14.36}$$

This expression generalizes (14.11) to more dimensions.

14.5 Delta function on the sphere

Up to this point we have analyzed the delta function in Cartesian coordinates. When the delta function is used in a coordinate system that is not Cartesian, additional terms appear in the definition of the delta function. This is illustrated in this section with the delta function on a sphere. Suppose we define a function $f(\theta, \varphi)$ on the unit sphere. Using definition (14.11) the action of the delta function on the sphere as expressed in the angles θ and φ can be written as

$$\int_0^{2\pi}\int_0^{\pi}\delta(\theta-\theta_0)\delta(\varphi-\varphi_0)f(\theta,\varphi)\,d\theta d\varphi=f(\theta_0,\varphi_0). \tag{14.37}$$

Every point on the unit sphere can be characterized by the unit vector $\hat{\mathbf{r}}$ that points from the origin to that point. It follows from the first identity of (4.7) that for given angles θ and φ this unit vector is given by

$$\hat{\mathbf{r}}=\begin{pmatrix}\sin\theta\cos\varphi\\ \sin\theta\sin\varphi\\ \cos\theta\end{pmatrix}; \tag{14.38}$$

a similar definition holds for the unit vector $\hat{\mathbf{r}}_0$ that corresponds to the angles θ_0 and φ_0.

In this section we rewrite (14.37) as an integration over the unit sphere. According to (4.35), the surface element on the sphere is given by $dS=r^2\sin\theta d\theta d\varphi$. On the unit sphere the radius is, by definition, given by $r=1$, so that on the unit sphere $dS=\sin\theta d\theta d\varphi$.

Problem a Show that (14.37) can be written as

$$\oint\frac{1}{\sin\theta}\delta(\theta-\theta_0)\delta(\varphi-\varphi_0)f(\theta,\varphi)\,dS=f(\theta_0,\varphi_0), \tag{14.39}$$

where the symbol $\oint\cdots dS$ denotes the integration over the unit sphere.

We can consider the function f to be a function of the angles θ and φ, but we can also see f as a function of the unit vector $\hat{\mathbf{r}}$, which means that $f=f(\hat{\mathbf{r}})$. Therefore we can write the previous expression as

$$\oint\frac{1}{\sin\theta}\delta(\theta-\theta_0)\delta(\varphi-\varphi_0)f(\hat{\mathbf{r}})\,dS=f(\hat{\mathbf{r}}_0). \tag{14.40}$$

Formally this can also be written as

$$\oint\delta(\hat{\mathbf{r}}-\hat{\mathbf{r}}_0)f(\hat{\mathbf{r}})\,dS=f(\hat{\mathbf{r}}_0), \tag{14.41}$$

where $\delta(\hat{\mathbf{r}}-\hat{\mathbf{r}}_0)$ defines the delta function on the unit sphere.

Problem b Show that

$$\delta(\hat{\mathbf{r}}-\hat{\mathbf{r}}_0)=\frac{1}{\sin\theta}\delta(\theta-\theta_0)\delta(\varphi-\varphi_0). \tag{14.42}$$

This expression shows that when the delta function is computed in a non-Cartesian coordinate system (such as spherical coordinates defined on the unit sphere) additional terms appear in the delta function that account for the curvilinear character of the coordinate system. The terms that appear in the delta function compensate the terms in the Jacobian that account for the curvilinear character of the employed coordinate system.

14.6 Self energy of the electron

Up to this point the delta function has been treated as a mathematical tool. However, it is often used as a description of the concept of a point charge or point mass. Physically, the idea of a point mass is that a finite mass M is concentrated at a certain point $\mathbf{r}_0$. The associated mass-density $\rho(\mathbf{r})$ is then equal to zero everywhere except at $\mathbf{r} = \mathbf{r}_0$. The integral of the mass-density must be equal to the total mass: $\int \rho(\mathbf{r}) d^3\mathbf{r} = M$, where $\int \cdots d^3\mathbf{r}$ denotes the three-dimensional volume integral.

Problem a Verify that these properties are satisfied by the mass-density

$$\rho(\mathbf{r}) = M\delta(\mathbf{r} - \mathbf{r}_0). \tag{14.43}$$

For the moment we restrict our attention to the concept of a point mass and the associated gravitational field, but because of the equivalence of the laws for the gravitational field and the electrostatic field the results can be applied to a stationary electric field as well.

It was shown in Section 8.2 that the gravitational field generated by a spherically symmetric body depends outside the body on the total mass only and not on the mass distribution within the body. For a mass centered at the origin ($\mathbf{r}_0 = 0$), the gravitational field is given by

$$\mathbf{g}(\mathbf{r}) = -\frac{GM}{r^2}\hat{\mathbf{r}} . \tag{8.5}$$

This gravitational field is associated with a gravitational potential $V(\mathbf{r})$ that is related to the gravity field by the expression

$$\mathbf{g}(\mathbf{r}) = -\nabla V(\mathbf{r}). \tag{14.44}$$

Problem b Use (8.5) to show that the gravitational potential for a point mass M located in the origin is given by

$$V(\mathbf{r}) = -\frac{GM}{r}. \tag{14.45}$$

(In the integration you encounter one integration constant. This integration constant follows from the requirement that the potential energy vanishes at infinity.)

In the example in this section the point mass serves as the source of the gravitational field. The response of any linear system to a source function plays an important role in mathematical physics because a general source of the field can always be written as a superposition of delta functions. The response to a delta function excitation is called the *Green's function*: this concept is treated in Chapters 18 and 19. We have seen in Section 14.1 that the delta function is singular, and it is in fact so singular that it cannot be considered to be a function. Very often, the response to such a singular source function is also singular at the location of that source function. This means that the Green's function is usually singular at the point of excitation.

At this point we have computed the gravitational field and its potential energy for a point mass, and it appears that these can be computed without any problems. However, there is a complication. As shown in expression (1.53) of Jackson [53] the energy E of a charge-density $\rho(\mathbf{r})$ that is placed in a potential $V(\mathbf{r})$ is given by

$$E = \frac{1}{2}\int \rho(\mathbf{r})V(\mathbf{r})\,d^3\mathbf{r}. \tag{14.46}$$

(In reference [53] this is derived for electrostatic energy, but the same expression holds for gravitational energy.)

Problem c Use (14.43) and (14.45) to show that the gravitational energy of a point charge placed in the origin is infinite.

This means that a point mass has infinite energy, which indicates that the concept of a point mass is not without complications when the energy is concerned.

Let us consider what the energy is of a spherically symmetric mass distribution when the mass M is homogeneously distributed in a sphere of radius R. The concept of the boxcar function as defined in (14.7) can easily be generalized to three dimensions by the following definition:

$$B_R(\mathbf{r}) \equiv \begin{cases} \dfrac{3}{4\pi R^3} & \text{for} \quad |\mathbf{r}| \le R \\ 0 & \text{for} \quad |\mathbf{r}| > R \end{cases} . \tag{14.47}$$

Problem d Show that $\int B_R(\mathbf{r})\, d^3\mathbf{r} = 1$ and use this to show that the mass-density of such a mass distribution is given by

$$\rho(\mathbf{r}) = M B_R(\mathbf{r}). \tag{14.48}$$

Outside the mass, the gravitational field is given by (8.5) and the associated potential energy is derived in (14.45). Inside the mass the gravitational field is given by (8.7).

Problem e Use (8.7), (14.44), and the requirement that the potential is continuous everywhere to derive that the potential energy is given by

$$V(\mathbf{r}) \equiv \begin{cases} -\dfrac{GM}{2R^3}(3R^2 - r^2) & \text{for} \quad |\mathbf{r}| \leq R \\ -\dfrac{GM}{r} & \text{for} \quad |\mathbf{r}| > R \end{cases}. \tag{14.49}$$

Problem f Use this expression and (14.46) to show that the gravitational energy is given by

$$E = -\frac{3GM^2}{5R}. \tag{14.50}$$

Note that when the radius R of the mass goes to zero, the gravitational energy becomes infinite; this is what we derived in problem c.

Let us consider this homogeneous mass distribution for the moment as a simplified model of the mass distribution of the Earth. It follows from (14.50) that the gravitational energy decreases without bound when the radius of the Earth is decreased. Since physical systems tend to minimize their energy, we can raise the question of why the radius of the Earth has not become smaller than its present value of about 6370 km. The fact that the potential energy is negative corresponds to the fact that mass always attracts itself. The decrease of the gravitational energy with decreasing radius thus corresponds to a gravitational collapse of the body. So we can phrase our question in a different way: why does the Earth not collapse to a black hole?

There are two effects that need to be considered. First, the gravitational force in the Earth leads to a compression of the material within the Earth. This compression induces elastic forces within the Earth. The radius of the Earth is dictated by the balance between the gravitational force and the elastic reactive force. If energy were used to describe this balance, one would state that the positive potential energy of the elastic deformation balances the unbounded negative growth of the gravitational energy as the radius is decreased. The second factor to take into account is that the gravitational fields that are treated in this section follow from Newton's law of gravitation. However, this law only holds for weak fields, and for very strong fields it should be replaced by the laws of general relativity [76]. It follows from (8.5) that the gravitational field of a point mass grows without bounds as one moves closer to the point mass. This means that ultimately Newton's law of gravitation ceases to be a good description of the field close to the point mass. This implies that although the concept of a point mass as described by a delta function is appealing, it is physically not without complications.

You may think that this is a purely academic issue. However, the same reasoning applies to electric point charges. By analogy with (14.50), it follows that the energy as the radius R goes to zero of a homogeneous charge distribution within a sphere of radius R and a total charge q is given by

$$E = + \frac{3q^2}{20\pi\varepsilon_0 R}. \tag{14.51}$$

The energy becomes infinite as the radius R tends to zero. Note that the minus sign in (14.50) is replaced by a plus sign. This corresponds to the fact that equal masses attract each other while equal charges repel each other. Let us now consider the electron as a homogeneous charge q with radius R. Expression (14.51) then states that an electron has an infinite energy.

Problem g Show that the energy of the electron is minimized when the radius R goes to infinity.

This means that an electron would grow beyond bounds in order to minimize its energy. Just as with the previous discussion of the gravitational energy of the Earth there are a number of reasons why this is not a physically accurate description. By analogy with the elastic forces in the Earth, one could argue there may be other forces acting within the electron that keep the charge together. However, the concept of an electron as a homogeneous charge distribution is physically not accurate. The charge of an electron is quantized, and the laws of electrostatics are not applicable to the description of the internal structure of the electron.

Nevertheless, the self energy of the electron is a long standing problem [117] because the quantum theory of the electron also predicts an infinite energy of the electron. The more advanced quantum field theory of the electron also predicts an infinite self energy of the electron, but in this case the singularity is logarithmic rather than algebraic [117]. This issue has both mathematical and philosophical aspects. Since the electron has a fixed quantized charge, one cannot consider an electron separately from its field. In fact, through quantum fluctuations the electron can interact with virtual particles that are generated in its field. This means that the energy of an electron that we observe is always a mixture of the energy the electron would have in the absence of its interaction with its field and the interaction with its field (and all virtual particles that may be generated in that field). This has led to the concept of renormalization [91, 125] in which one accounts for the fact that the energy of the electron in the absence of electromagnetic fields is different from the mass that we observe.

15

Fourier analysis

Fourier analysis is concerned with the decomposition of signals into sine and cosine waves. This technique is of obvious relevance for spectral analysis where one decomposes a time signal into its different frequency components. As an example, the spectrum of a low-C on a soprano saxophone is shown in Figure 15.1. However, the use of Fourier analysis goes far beyond this application because Fourier analysis can also be used to find solutions of differential equations and a large number of other applications. In this chapter the real Fourier transform on a finite interval is used as a starting point. From this the complex Fourier transform and the Fourier transform on an infinite interval are derived. At several stages of the analysis, the similarity between Fourier analysis and linear algebra is treated.

15.1 Real Fourier series on a finite interval

Consider a function $f(x)$ that is defined on the interval $-L < x \leq L$. This interval is of length $2L$, and let us assume that $f(x)$ is periodic with period $2L$. This means that if one translates this function over a distance $2L$ the value does not change:

$$f(x + 2L) = f(x). \tag{15.1}$$

We want to expand this function into a set of basis functions. Since $f(x)$ is periodic with period $2L$, these basis functions must be periodic with the same period.

Problem a Show that the functions $\cos(n\pi x/L)$ and $\sin(n\pi x/L)$ with integer n are periodic with period $2L$, that is show that these functions satisfy (15.1).

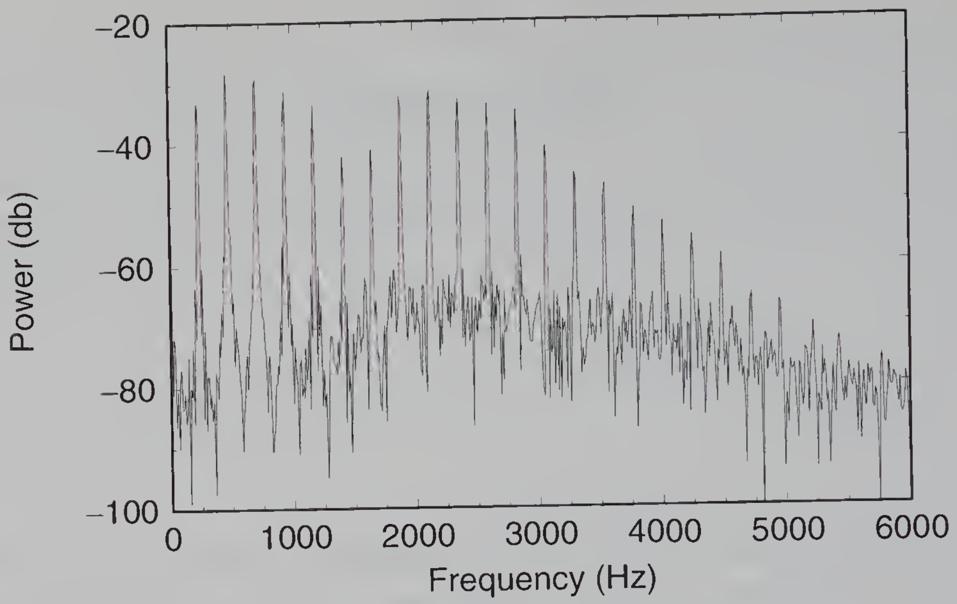

Fig. 15.1 The energy of the sound made by the author playing a low-C on his soprano saxophone as a function of frequency. The unit used for the horizontal axis is the hertz (the number of oscillations per second), the unit on the vertical axis is decibels (a logarithmic measure of energy).

The main statement of Fourier analysis is that one can write $f(x)$ as a superposition of these periodic sine and cosine waves:

$$f(x) = \frac{1}{2}a_0 + \sum_{n=1}^{\infty} a_n \cos(n\pi x/L) + \sum_{n=1}^{\infty} b_n \sin(n\pi x/L). \qquad (15.2)$$

The factor 1/2 with the coefficient a_0 has no special significance, and is used in order to simplify subsequent expressions. To show that (15.2) is actually true is not trivial. Providing this proof essentially amounts to showing that the functions $\cos(n\pi x/L)$ and $\sin(n\pi x/L)$ actually contain enough "degrees of freedom" to describe $f(x)$. However, since $f(x)$ is a function of a continuous variable x this function has infinitely many degrees of freedom and since there are infinitely many coefficients a_n and b_n counting the number of degrees of freedom does not work. Mathematically one would say that one needs to show that the set of functions $\cos(n\pi x/L)$ and $\sin(n\pi x/L)$ is a "complete set." We will not concern ourselves here with this proof, and simply start working with the Fourier series (15.2).

At this point it is not yet clear what the coefficients a_n and b_n are. In order to derive these coefficients one needs to use the following integrals:

$$\int_{-L}^{L} \cos^2(n\pi x/L)\,dx = \int_{-L}^{L} \sin^2(n\pi x/L)\,dx = L \qquad (n \geq 1); \qquad (15.3)$$

$$\int_{-L}^{L} \cos(n\pi x/L)\cos(m\pi x/L)\,dx = 0 \qquad \text{if} \quad n \neq m; \qquad (15.4)$$

$$\int_{-L}^{L} \sin(n\pi x/L)\sin(m\pi x/L)\,dx = 0 \qquad \text{if} \quad n \neq m; \tag{15.5}$$

$$\int_{-L}^{L} \cos(n\pi x/L)\sin(m\pi x/L)\,dx = 0 \qquad \text{all} \quad n, m. \tag{15.6}$$

Problem b Derive these identities. In doing so you need to use trigonometric identities such as $\cos\alpha\cos\beta = [\cos(\alpha+\beta)+\cos(\alpha-\beta)]/2$. If you have difficulty deriving these identities you can consult a textbook such as Boas [19].

Problem c In order to find the coefficient b_m, multiply the Fourier expansion (15.2) by $\sin(m\pi x/L)$, integrate the result from $-L$ to L and use (15.3)–(15.6) to evaluate the integrals. Show that this gives:

$$b_n = \frac{1}{L}\int_{-L}^{L} f(x)\sin(n\pi x/L)\,dx. \tag{15.7}$$

Problem d Use a similar analysis to show that:

$$a_n = \frac{1}{L}\int_{-L}^{L} f(x)\cos(n\pi x/L)\,dx. \tag{15.8}$$

In deriving this result treat the cases $n \neq 0$ and $n = 0$ separately. It is now clear why the factor $1/2$ was introduced in the a_0 term of (15.2); without this factor expression (15.8) would have an additional factor 2 for $n = 0$.

There is a close relation between the Fourier series (15.2) and the coefficients given in the expressions above and the projection of a vector on a number of basis vectors in linear algebra as shown in Section 13.1. To see this relation we restrict ourselves for simplicity to functions $f(x)$ that are odd functions of x: $f(-x) = -f(x)$, but this restriction is by no means essential. For these functions all coefficients a_n are equal to zero. As an analog of a basis vector in linear algebra let us define the following basis function $u_n(x)$:

$$u_n(x) \equiv \frac{1}{\sqrt{L}}\sin(n\pi x/L). \tag{15.9}$$

An essential ingredient in the projection operators of Section 13.1 is the inner product between vectors. It is also possible to define an inner product for functions, and for the present example the inner product of two functions $f(x)$ and $g(x)$ is defined as:

$$(f\cdot g) \equiv \int_{-L}^{L} f(x)g(x)\,dx. \tag{15.10}$$

Problem e The basis functions $u_n(x)$ defined in (15.9) are the analog of a set of orthonormal basis vectors. To see this, use (15.3) and (15.5) to show that

$$(u_n \cdot u_m) = \delta_{nm}, \tag{15.11}$$

where δ_{nm} is the Kronecker delta.

This expression implies that the basis functions $u_n(x)$ are mutually orthogonal. If the norm of such a basis function is defined as $\|u_n\| \equiv \sqrt{(u_n \cdot u_n)}$, (15.11) implies that the basis functions are normalized (i.e. have norm 1). These functions are the generalization of orthogonal unit vectors to a function space. The (odd) function $f(x)$ can be written as a sum of the basis functions $u_n(x)$:

$$f(x) = \sum_{n=1}^{\infty} c_n u_n(x). \tag{15.12}$$

Problem f Take the inner product of (15.12) with $u_m(x)$ and show that $c_m = (u_m \cdot f)$. Use this to show that the Fourier expansion of $f(x)$ can be written as $f(x) = \sum_{n=1}^{\infty} u_n(x)\,(u_n \cdot f)$, and that on leaving out the explicit dependence on the variable x the result is

$$f = \sum_{n=1}^{\infty} u_n\,(u_n \cdot f). \tag{15.13}$$

This equation bears a close resemblance to the expression derived in Section 13.1 for the projection of vectors. The projection of a vector **v** along a unit vector $\hat{\mathbf{n}}$ was shown to be

$$\mathbf{P}\mathbf{v} = \hat{\mathbf{n}}\,(\hat{\mathbf{n}} \cdot \mathbf{v}). \tag{13.2}$$

A comparison with (15.13) shows that $u_n(x)\,(u_n \cdot f)$ can be interpreted as the projection of the function $f(x)$ on the function $u_n(x)$. To reconstruct the function, one must sum over the projections along all basis functions, hence the summation in (15.13). It is shown in (13.12) that in order to find the projection of the vector **v** onto the subspace spanned by a finite number of orthonormal basis vectors one simply has to sum the projections of the vector **v** on all the basis vectors that span the subspace: $\mathbf{P}\mathbf{v} = \sum_i \hat{\mathbf{n}}_i\,(\hat{\mathbf{n}}_i \cdot \mathbf{v})$. In a similar way, one can sum the Fourier series (15.13) over only a limited number of basis functions to obtain the projection of $f(x)$ on a limited number of basis functions:

$$f_{\mathit{filtered}} = \sum_{n=n_1}^{n_2} u_n\,(u_n \cdot f). \tag{15.14}$$

In this expression it was assumed that only values $n_1 \leq n \leq n_2$ have been used. The projected function is called $f_{filtered}$ because this projection is a filtering operation.

Problem g Show that the functions $u_n(x)$ are sinusoidal waves with wavelength $\lambda = 2L/n$.

This means that restricting the n-values in the sum (15.14) amounts to using only wavelengths between $2L/n_2$ and $2L/n_1$ for the projected function. Since only certain wavelengths are used, this projection really acts as a filter that allows only certain wavelengths in the filtered function.

It is the filtering property that makes the Fourier transform so useful for filtering data sets by excluding wavelengths that are unwanted. In fact, the Fourier transform forms the basis of digital filtering techniques that have many applications in science and engineering, see for example the books of Claerbout [27] and Robinson and Treitel [90].

15.2 Complex Fourier series on a finite interval

In the theory of the preceding section there is no reason why the function $f(x)$ should be real. Although the basis functions $\cos(n\pi x/L)$ and $\sin(n\pi x/L)$ are real, the Fourier sum (15.2) can be complex because the coefficients a_n and b_n can be complex. The equation of de Moivre gives the relation between these basis functions and complex exponential functions:

$$e^{in\pi x/L} = \cos(n\pi x/L) + i\,\sin(n\pi x/L). \tag{15.15}$$

This expression can be used to rewrite the Fourier series (15.2) using the basis functions $e^{in\pi x/L}$ rather than sines and cosines.

Problem a Replace n by $-n$ in (15.15) to show that:

$$\left.\begin{aligned} \cos(n\pi x/L) &= \frac{1}{2}\left(e^{in\pi x/L} + e^{-in\pi x/L}\right), \\ \sin(n\pi x/L) &= \frac{1}{2i}\left(e^{in\pi x/L} - e^{-in\pi x/L}\right). \end{aligned}\right\} \tag{15.16}$$

Problem b Insert this relation in the Fourier series (15.2) to show that this Fourier series can also be written as:

$$f(x) = \sum_{n=-\infty}^{\infty} c_n\, e^{in\pi x/L}, \tag{15.17}$$

with the coefficients c_n given by:

$$\left.\begin{aligned} c_n &= (a_n - ib_n)/2 && \text{for} \quad n > 0, \\ c_n &= (a_{|n|} + ib_{|n|})/2 && \text{for} \quad n < 0, \\ c_0 &= a_0/2. \end{aligned}\right\} \tag{15.18}$$

Note that the absolute value $|n|$ is used for $n < 0$.

Problem c Explain why the n-summation in (15.17) extends from $-\infty$ to ∞ rather than from 0 to ∞.

Problem d Relations (15.7) and (15.8) can be used to express the coefficients c_n in the function $f(x)$. Treat the cases $n > 0$, $n < 0$ and $n = 0$ separately to show that for all values of n the coefficient c_n is given by:

$$c_n = \frac{1}{2L} \int_{-L}^{L} f(x) e^{-in\pi x/L} dx. \tag{15.19}$$

The sum (15.17) with (15.19) constitutes the complex Fourier transform over a finite interval. Again, there is a close analogy with the projections of vectors shown in Section 13.1. Before we can explore this analogy, the inner product between two complex functions $f(x)$ and $g(x)$ needs to be defined. This inner product is not given by $(f \cdot g) = \int f(x) g(x) dx$. The reason for this is that the length of a vector is defined by $\|\mathbf{v}\|^2 = (\mathbf{v} \cdot \mathbf{v})$, and a straightforward generalization of this expression to functions using the inner product given above would give for the norm of the function: $\|f\|^2 = (f \cdot f) = \int f^2(x) dx$. However, when $f(x)$ is purely imaginary this would lead to a negative norm. This can be avoided by defining the inner product of two complex functions by:

$$(f \cdot g) \equiv \int_{-L}^{L} f^*(x) g(x)\, dx, \tag{15.20}$$

where the asterisk denotes the complex conjugate.

Problem e Show that with this definition the norm of $f(x)$ is given by $\|f\|^2 = (f \cdot f) = \int |f(x)|^2 \, dx$.

With this inner product the norm of the function is guaranteed to be positive. Now that we have an inner product, the analogy with the projections in linear algebra can be explored. In order to do this, define the following basis functions:

$$u_n(x) \equiv \frac{1}{\sqrt{2L}} e^{in\pi x/L}. \tag{15.21}$$

Problem f Show that these functions are orthonormal with respect to the inner product (15.20), that is show that:

$$(u_n \cdot u_m) = \delta_{nm}. \tag{15.22}$$

Pay special attention to the normalization of these functions; that is to the case $n = m$.

Problem g Expand the function $f(x)$ in these basis functions, $f(x) = \sum_{n=-\infty}^{\infty} \gamma_n u_n(x)$ and show that $f(x)$ can be written as:

$$f = \sum_{n=-\infty}^{\infty} u_n \, (u_n \cdot f). \tag{15.23}$$

Problem h Make the comparison between this expression and the expressions for the projections of vectors in Section 13.1.

15.3 Fourier transform on an infinite interval

In several applications, one wants to compute the Fourier transform of a function that is defined on an infinite interval. This amounts to taking the limit $L \to \infty$. However, a simple inspection of (15.19) shows that one cannot simply take the limit $L \to \infty$ of the expressions of the previous section because in that limit c_n is a possible infinite integral divided by infinity, which is poorly defined. In order to define the Fourier transform for an infinite interval define the variable k by:

$$k \equiv \frac{n\pi}{L}. \tag{15.24}$$

An increment Δn corresponds to an increment Δk given by $\Delta k = \pi \Delta n / L$. In the summation over n in the Fourier expansion (15.17), n is incremented by unity: $\Delta n = 1$. This corresponds to an increment $\Delta k = \pi / L$ of the variable k. In the limit $L \to \infty$ this increment goes to zero, which implies that the summation over n should be replaced by an integration over k:

$$\sum_{n=-\infty}^{\infty} (\cdots) \to \frac{\Delta n}{\Delta k} \int_{-\infty}^{\infty} (\cdots) \, dk = \frac{L}{\pi} \int_{-\infty}^{\infty} (\cdots) \, dk \qquad \text{as } L \to \infty. \tag{15.25}$$

Problem a Explain the presence of the factor $\Delta n / \Delta k$ and prove the last identity.

This is not enough to generalize the Fourier transform of the previous section to an infinite interval. As noted earlier, the coefficients c_n are poorly defined in the limit

$L \to \infty$. Also note that the integral on the right-hand side of (15.25) is multiplied by L/π, and this coefficient is infinite in the limit $L \to \infty$. Both complications can be solved by defining the following function:

$$F(k) \equiv \frac{L}{\pi} c_n , \tag{15.26}$$

where the relation between k and n is given by (15.24).

Problem b Show that with the replacements (15.25) and (15.26) the limit $L \to \infty$ of the complex Fourier transform (15.17) and (15.19) can be taken and that the result can be written as:

$$f(x) = \int_{-\infty}^{\infty} F(k) e^{ikx} dk , \tag{15.27}$$

$$F(k) = \frac{1}{2\pi} \int_{-\infty}^{\infty} f(x) e^{-ikx} dx . \tag{15.28}$$

15.4 Fourier transform and the delta function

In this section the Fourier transform of the delta function is treated. This is not only useful in a variety of applications, but it will also establish the relation between the Fourier transform and the closure relation introduced in Section 13.1. Consider the delta function centered at $x = x_0$:

$$f(x) = \delta(x - x_0) . \tag{15.29}$$

Problem a Show that the Fourier transform $F(k)$ of $f(x)$ as defined in expression (15.28) is given by:

$$F(k) = \frac{1}{2\pi} e^{-ikx_0} . \tag{15.30}$$

Problem b Show that this implies that the Fourier transform of the delta function $\delta(x)$ centered at $x = 0$ is a constant. Determine this constant.

Problem c Use (15.27) to show that

$$\delta(x - x_0) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{ik(x-x_0)} dk . \tag{15.31}$$

Problem d Use a similar analysis to derive that

$$\delta\left(k-k_0\right)=\frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-i(k-k_0)x}\,dx. \tag{15.32}$$

These expressions are useful in a number of applications. Again, there is a close analogy between these expressions and the projection of vectors introduced in Section 13.1. To establish this connection let us define the following basis functions:

$$u_k(x)\equiv\frac{1}{\sqrt{2\pi}}e^{ikx}, \tag{15.33}$$

and use the inner product defined in (15.20) with the integration limits extending from $-\infty$ to ∞.

Problem e Show that (15.32) implies that

$$\left(u_k\cdot u_{k_0}\right)=\delta(k-k_0). \tag{15.34}$$

This implies that the functions $u_k(x)$ form an orthonormal set, because this relation generalizes (15.22) to a continuous basis of functions.

Problem f Use (15.31) to derive that:

$$\int_{-\infty}^{\infty} u_k(x)u_k^*(x_0)\,dk=\delta\left(x-x_0\right). \tag{15.35}$$

This expression is the counterpart of the closure relation (13.13) introduced in Section 13.1 for finite-dimensional vector spaces. Note that the delta function $\delta\left(x-x_0\right)$ plays the role of the identity operator **I** with components $I_{ij}=\delta_{ij}$ in (13.13) and that the summation $\sum_{i=1}^{N}$ over the basis vectors is replaced by an integration $\int_{-\infty}^{\infty}dk$ over the basis functions. Both differences are due to the fact that we are dealing in this section with an infinite-dimensional function space rather than a finite-dimensional vector space. Also note that in (15.35) the complex conjugate of $u_k(x_0)$ is taken. The reason for this is that for complex unit vectors $\hat{\mathbf{n}}$ the transpose in the completeness relation (13.13) should be replaced by the Hermitian conjugate. This involves taking the complex conjugate as well as taking the transpose.

15.5 Changing the sign and scale factor

In the Fourier transformation (15.27) from the wavenumber domain (k) to the position domain (x), the exponent has a plus sign, e^{+ikx}, and the coefficient multiplying the integral is given by 1. In other texts on Fourier transforms you may encounter a different sign in the exponent and different scale factors are sometimes used. For example, the exponent in the Fourier transformation from the wavenumber domain to the position domain may have a minus sign, e^{-ikx}, and there may be a scale factor such as $1/\sqrt{2\pi}$. It turns out that there is a freedom in choosing the sign of the exponential as well as in the scaling of the Fourier transform. We first study the effect of a scaling parameter on the Fourier transform.

Problem a Let the function $F(k)$ defined in (15.28) be related to a new function $\tilde{F}(k)$ by a scaling with a scale factor $C : F(k) = C\tilde{F}(k)$. Use (15.27) and (15.28) to show that:

$$f(x) = C\int_{-\infty}^{\infty} \tilde{F}(k)e^{ikx}\,dk, \tag{15.36}$$

$$\tilde{F}(k) = \frac{1}{2\pi C}\int_{-\infty}^{\infty} f(x)e^{-ikx}\,dx. \tag{15.37}$$

These expressions are equivalent to the original Fourier transform pair (15.27) and (15.28). The constant C is completely arbitrary. This implies that one may take any multiplication constant for the Fourier transform; the only restriction is that the product of the coefficients for Fourier transform and the backward transform is equal to $1/2\pi$.

Problem b Prove this last statement.

In the literature, notably in quantum mechanics, one often encounters the Fourier transform pair defined by the value $C = 1/\sqrt{2\pi}$. This leads to the Fourier transform pair:

$$f(x) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} \tilde{F}(k)e^{ikx}\,dk, \tag{15.38}$$

$$\tilde{F}(k) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} f(x)e^{-ikx}\,dx. \tag{15.39}$$

This normalization not only has the advantage that the multiplication factors for the forward and backward transformations are identical ($1/\sqrt{2\pi}$), but the constants are also identical to the constant used in (15.33) to create a set of orthonormal functions.

Next we investigate a change in the sign of the exponent in the Fourier transform. To do this, we will use the function $\tilde{F}(k)$ defined by: $\tilde{F}(k) = F(-k)$.

Problem c Change the integration variable k in (15.27) to $-k$ and show that the Fourier transform pair (15.27) and (15.28) is equivalent to:

$$f(x) = \int_{-\infty}^{\infty} \tilde{F}(k) e^{-ikx} dk, \tag{15.40}$$

$$\tilde{F}(k) = \frac{1}{2\pi} \int_{-\infty}^{\infty} f(x) e^{ikx} dx. \tag{15.41}$$

Note that these expressions differ from the earlier expressions only by the sign of the exponents. This means that there is a freedom in the choice of this sign. It does not matter which sign convention you use. Any choice of the sign and the multiplication constant for the Fourier transform can be used as long as:

> *(i) The product of the constants for the forward and backward transform is equal to $1/2\pi$ and (ii) the sign of the exponent for the forward and the backward transformation is opposite.*

In this book, the Fourier transform pair (15.27) and (15.28) is generally used for the Fourier transform from the space (x) domain to the wavenumber (k) domain.

Of course, the Fourier transform can also be used to transform a function in the time (t) domain to the frequency (ω) domain. Perhaps illogically the following convention is used in this book for this Fourier transform pair:

$$f(t) = \int_{-\infty}^{\infty} F(\omega) e^{-i\omega t} d\omega, \tag{15.42}$$

$$F(\omega) = \frac{1}{2\pi} \int_{-\infty}^{\infty} f(t) e^{i\omega t} dt. \tag{15.43}$$

The reason for this choice is that the combined Fourier transform from the (x, t)-domain to the (k, ω)-domain that is obtained by combining (15.27) and (15.42) is given by:

$$f(x, t) = \iint_{-\infty}^{\infty} F(k, \omega) e^{i(kx - \omega t)} dk d\omega. \tag{15.44}$$

The function $e^{i(kx-\omega t)}$ in this integral describes a wave that moves for positive values of k and ω in the direction of increasing values of x. To see this, let us assume that we are at a crest of this wave and that we follow the motion of the crest over a time Δt and we want to find the distance Δx that the crest has moved in that time interval. If we follow a wave crest, the phase of the wave is constant, and hence $kx - \omega t$ is constant.

Problem d Show that this implies that $\Delta x = c\Delta t$, with c given by $c = \omega/k$. Why does this imply that the wave moves with velocity c?

The exponential in the double Fourier transform (15.44) therefore describes for positive values of ω and k a wave traveling in the positive direction with velocity $c = \omega/k$. However, note that this is no proof that we should use the Fourier transform (15.44) and not a transform with a different choice of the sign in the exponent. In fact, one should realize that in the Fourier transform (15.44) one needs to integrate over *all* values of ω and k so that negative values of ω and k contribute to the integral as well.

Problem e Use (15.28) and (15.43) to derive the inverse of the double Fourier transform (15.44).

15.6 Convolution and correlation of two signals

There are different ways in which one can combine signals to create a new signal. In this section the convolution and correlation of two signals is treated. For the sake of argument the signals are taken to be functions of time, and the Fourier transform pair (15.42) and (15.43) is used for the forward and inverse Fourier transforms. Suppose a function $f(t)$ has a Fourier transform $F(\omega)$ defined by (15.42) and another function $h(t)$ has a similar Fourier transform $H(\omega)$:

$$h(t) = \int_{-\infty}^{\infty} H(\omega)e^{-i\omega t}d\omega. \tag{15.45}$$

The two Fourier transforms $F(\omega)$ and $H(\omega)$ can be multiplied in the frequency domain, and we want to find out what the Fourier transform of the product $F(\omega)H(\omega)$ is in the time domain.

Problem a Show that:

$$F(\omega)H(\omega) = \frac{1}{(2\pi)^2}\iint_{-\infty}^{\infty} f(t_1)h(t_2)e^{i\omega(t_1+t_2)}dt_1dt_2. \tag{15.46}$$

Problem b Show that after a Fourier transformation this function corresponds in the time domain to:

$$\int_{-\infty}^{\infty} F(\omega)H(\omega)e^{-i\omega t}d\omega = \frac{1}{(2\pi)^2}\iiint_{-\infty}^{\infty} f(t_1)h(t_2)e^{i\omega(t_1+t_2-t)}dt_1dt_2d\omega. \tag{15.47}$$

Problem c Use the representation (15.31) of the delta function to carry out the integration over ω and show that this gives:

$$\int_{-\infty}^{\infty} F(\omega)H(\omega)e^{-i\omega t}\,d\omega = \frac{1}{2\pi}\iint_{-\infty}^{\infty} f(t_1)h(t_2)\delta(t_1+t_2-t)\,dt_1dt_2. \quad (15.48)$$

Problem d The integration over t_1 can now be carried out. Do this, and show that after renaming the variable t_2 as τ the result can be written as:

$$\int_{-\infty}^{\infty} F(\omega)H(\omega)e^{-i\omega t}\,d\omega = \frac{1}{2\pi}\int_{-\infty}^{\infty} f(t-\tau)h(\tau)\,d\tau = \frac{1}{2\pi}(f*h)(t). \quad (15.49)$$

The τ-integral in the middle term is called the *convolution* of the functions f and h; this operation is denoted by the symbol $(f*h)$. Equation (15.49) states that a multiplication of the spectra of two functions in the frequency domain corresponds to the convolution of these functions in the time domain. For this reason, (15.49) is called the *convolution theorem*. This theorem is schematically indicated in the following diagram:

$$\begin{aligned} f(t) &\longleftrightarrow F(\omega), \\ h(t) &\longleftrightarrow H(\omega), \\ \frac{1}{2\pi}(f*h) &\longleftrightarrow F(\omega)H(\omega). \end{aligned}$$

Note that in the convolution theorem, a scale factor $1/2\pi$ is present on the left-hand side. This scale factor depends on the choice of the scale factors that one uses in the Fourier transform, see Section 15.5.

Problem e Use a change of the integration variable to show that the convolution of f and h can be written in the following two ways:

$$(f*h)(t) = \int_{-\infty}^{\infty} f(t-\tau)h(\tau)\,d\tau = \int_{-\infty}^{\infty} f(\tau)h(t-\tau)\,d\tau. \quad (15.50)$$

Problem f In order to see what the convolution theorem looks like when a different scale factor is used in the Fourier transform, define $F(\omega) = C\tilde{F}(\omega)$, and a similar scaling for $H(\omega)$. Show that with this choice of the scale factors, the Fourier transform of $\tilde{F}(\omega)\tilde{H}(\omega)$ is in the time domain given by $(1/2\pi C)(f*h)(t)$. Hint: first determine the scale factor that one needs to use in the transformation from the frequency domain to the time domain.

The convolution of two time series plays an important role in exploration geophysics. Suppose one carries out a seismic experiment in which one uses a source such as dynamite or a vibrator to generate waves that propagate through the Earth. Let the source signal in the frequency domain be given by $S(\omega)$. The waves reflect at layers in the Earth and are recorded by geophones. In the ideal case, the source signal would have the shape of a simple spike, and the waves reflected by all the reflectors would show up as a sequence of individual spikes. In that case the recorded data would indicate the true reflectors in the Earth. Let the signal $r(t)$ recorded in this ideal case have a Fourier transform $R(\omega)$ in the frequency domain. The problem that one faces is that a real seismic source is often not very impulsive. If the recorded data $d(t)$ have a Fourier transform $D(\omega)$ in the frequency domain, then this Fourier transform is given by

$$D(\omega) = R(\omega)S(\omega). \tag{15.51}$$

One is only interested in $R(\omega)$ which is the Earth response in the frequency domain, but in practice one records the product $R(\omega)S(\omega)$. In the time domain this is equivalent to saying that one has recorded the convolution $\int_{-\infty}^{\infty} r(\tau)s(t-\tau)d\tau$ of the Earth response with the source signal, but that one is only interested in the Earth response $r(t)$. One would like to "undo" this convolution; this process is called *deconvolution*. Carrying out the deconvolution seems trivial in the frequency domain. According to (15.51) one only needs to divide the data in the frequency domain by the source spectrum $S(\omega)$ to obtain $R(\omega)$. The problem is that in practice one often does not know the source spectrum $S(\omega)$. This makes seismic deconvolution a difficult process; see the collection of articles compiled by Webster [116]. It has been strongly argued by Ziolkowski [126] that the seismic industry should make a larger effort to record the source signal accurately.

The convolution of two signals was obtained in this section by taking the product $F(\omega)H(\omega)$ and carrying out a Fourier transformation back to the time domain. The same steps can be taken by multiplying $F(\omega)$ by the complex conjugate $H^*(\omega)$ and applying a Fourier transformation to go to the time domain.

Problem g Take steps similar to those in the derivation of the convolution to show that

$$\int_{-\infty}^{\infty} F(\omega)H^*(\omega)e^{-i\omega t}\,d\omega = \frac{1}{2\pi}\int_{-\infty}^{\infty} f(t+\tau)h^*(\tau)\,d\tau. \tag{15.52}$$

The right-hand side of this expression is called the *correlation* of the functions $f(t)$ and $h^*(t)$. Note that this expression is similar to the convolution theorem (15.50). This result implies that the Fourier transform of the product of a function and the complex conjugate in the frequency domain corresponds to the correlation in the time domain. Note again the constant $1/2\pi$ on the right-hand side. This constant again depends on the scale factors used in the Fourier transform.

Problem h Set $t = 0$ in (15.52) and let the function $h(t)$ be equal to $f(t)$. Show that this gives:

$$\int_{-\infty}^{\infty} |F(\omega)|^2 \, d\omega = \frac{1}{2\pi} \int_{-\infty}^{\infty} |f(t)|^2 \, dt. \tag{15.53}$$

This is known as *Parseval's theorem.* To see its significance, note that $\int_{-\infty}^{\infty} |f(t)|^2 \, dt = (f \cdot f)$, with the inner product of (15.20) with t as the integration variable and with the integration extending from $-\infty$ to ∞. Since $\sqrt{(f \cdot f)}$ is the norm of f measured in the time domain, and since $\int_{-\infty}^{\infty} |F(\omega)|^2 \, d\omega$ is the square of the norm of F measured in the frequency domain, Parseval's theorem states that with this definition of the norm, the norm of a function is equal in the time domain and is in the frequency domain (up to the scale factor $1/2\pi$).

15.7 Linear filters and the convolution theorem

Let us consider a linear system that has an output signal $o(t)$ when it is given an input signal $i(t)$, see Figure 15.2. There are numerous examples of this kind of system. As an example, consider a damped harmonic oscillator that is driven by a force; this system is described by the differential equation $\ddot{x} + 2\beta\dot{x} + \omega_0^2 x = F/m$, where the dots denotes a time derivative. The force $F(t)$ can be seen as the input signal, and the response $x(t)$ of the oscillator can be seen as the output signal. The relation between the input signal and the output signal is governed by the characteristics of the system under consideration; in this example it is the physics of the damped harmonic oscillator that determines the relation between the input signal $F(t)$ and the output signal $x(t)$.

Note that we have not defined yet what a *linear* filter is. A filter is linear when an input $c_1 i_1(t) + c_2 i_2(t)$ leads to an output $c_1 o_1(t) + c_2 o_2(t)$, in which $o_1(t)$ is the output corresponding to the input $i_1(t)$ and $o_2(t)$ is the output corresponding to the input $i_2(t)$.

Problem a Can you think of another example of a linear filter?

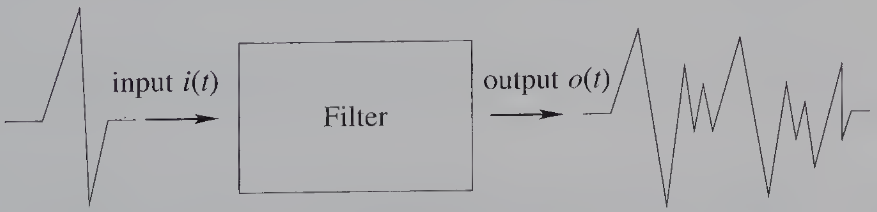

Fig. 15.2 Schematic representation of a linear filter.

Problem b Can you think of a system that has one input signal and one output signal, in which these signals are related through a nonlinear relation? This would be an example of a nonlinear filter; the theory of this section would not apply to such a filter.

It is possible to determine the output $o(t)$ for any input $i(t)$ if the output to a delta function input is known. Consider the special input signal $\delta(t-\tau)$ that consists of a delta function centered at $t=\tau$. Since a delta function has "zero-width" (if it has a width at all) such an input function is impulsive. Let the output for this particular input be denoted by $g(t,\tau)$. Since this function is the response at time t to an impulsive input at time τ this function is called the *impulse response*:

> *The impulse response function $g(t,\tau)$ is the output of the system at time t due to an impulsive input at time τ.*

How can the impulse response be used to find the response to an arbitrary input function? Any input function can be written as:

$$i(t)=\int_{-\infty}^{\infty}\delta(t-\tau)i(\tau)\,d\tau. \tag{15.54}$$

This identity follows from the definition of the delta function. However, we can also look at this expression from a different point of view. The integral on the right-hand side of (15.54) can be seen as a superposition of infinitely many delta functions $\delta(t-\tau)$. Each delta function when considered as a function of t is centered at time τ. Since we integrate over τ these different delta functions are superposed to construct the input signal $i(t)$. Each of the delta functions in the integral (15.54) is multiplied by $i(\tau)$. This term plays the role of a coefficient that gives a weight to the delta function $\delta(t-\tau)$.

At this point it is crucial to use that the filter is linear. Since the response to the input $\delta(t-\tau)$ is the impulse response $g(t,\tau)$, and since the input can be written as the superposition (15.54) of delta function input signals $\delta(t-\tau)$, the output can be written as the same superposition of impulse response signals $g(t,\tau)$:

$$o(t)=\int_{-\infty}^{\infty}g(t,\tau)i(\tau)\,d\tau. \tag{15.55}$$

Problem c Carefully compare (15.54) and (15.55). Note the similarity and make sure you understand the reasoning that has led to (15.55).

You may find this "derivation" of (15.55) rather vague. The notion of the impulse response will be treated in much greater detail in Chapter 18 because it plays a crucial role in mathematical physics.

At this point we make another assumption about the system. Apart from its linearity we will also assume it is *invariant for translations in time* . This is a complex way of saying that we assume that the properties of the filter do not change with time. This is the case for the damped harmonic oscillator described at the beginning of this section. However, this oscillator would not be invariant for translations in time if the damping parameter were a function of time as well: $\beta = \beta(t)$. In that case, the system would give different responses when the same input is used at different times.

When the properties of the filter do not depend on time, the impulse response $g(t, \tau)$ depends only on the *difference* $t - \tau$. To see this, consider the damped harmonic oscillator again. The response at a certain time depends only on the time that has lapsed between the excitation at time τ and the time of observation t. Therefore, for a time-invariant filter:

$$g(t, \tau) = g(t - \tau). \tag{15.56}$$

Inserting this in (15.55) shows that for a linear time-invariant filter the output is given by the convolution of the input with the impulse response:

$$o(t) = \int_{-\infty}^{\infty} g(t - \tau) i(\tau)\, d\tau = (g * i)(t). \tag{15.57}$$

Problem d Let the Fourier transform of $i(t)$ be given by $I(\omega)$, the Fourier transform of $o(t)$ by $O(\omega)$ and the Fourier transform of $g(t)$ by $G(\omega)$. Use (15.57) to show that these Fourier transforms are related by:

$$O(\omega) = 2\pi G(\omega) I(\omega). \tag{15.58}$$

Expressions (15.57) and (15.58) are key results in the theory in linear time-invariant filters. The first of these states that one only needs to know the response $g(t)$ to a single impulse to compute the output of the filter to *any* input signal $i(t)$. Equation (15.58) has two important consequences. First, if one knows the Fourier transform $G(\omega)$ of the impulse response, one can compute the Fourier transform $O(\omega)$ of the output. An inverse Fourier transform then gives the output $o(t)$ in the time domain.

Problem e Show that for a fixed value of ω that $G(\omega)e^{-i\omega t}$ is the response of the system to the input signal $e^{-i\omega t}$.

This means that if one knows the response of the filter to the harmonic signal $e^{-i\omega t}$ at any frequency, one knows $G(\omega)$ and the response to any input signal can be determined.

The second important consequence of (15.58) is that the output at frequency ω depends only on the input and impulse response at the same frequency ω, and not on other frequencies. This last property does not hold for nonlinear systems, because for these different frequencies components of the input signal are mixed by the nonlinearity of the system. An example of this phenomenon is the Earth's climate which has variations which contain frequency components that cannot be explained by periodic variations in the orbital parameters in the Earth, but which are due to the nonlinear character of the climate response to the amount of energy received by the Sun [96].

The fact that a filter can be used either by specifying its Fourier transform $G(\omega)$ (or equivalently the response to an harmonic input $e^{-i\omega t}$) or by prescribing the impulse response $g(t)$ implies that a filter can be designed either in the frequency domain or in the time domain. In Section 15.8 the action of a filter in the time domain is described. A Fourier transform then leads to a compact description of the filter response in the frequency domain. In Section 15.9 the converse route is taken; the filter is designed in the frequency domain, and a Fourier transform is used to derive an expression for the filter in the time domain.

As a last reminder it should be mentioned that although the theory of linear filters is introduced here for filters that act in the time domain, the theory is equally valid for filters in the spatial domain. In the latter case, the wavenumber k plays the role that the angular frequency played in this section. Since there may be more than one spatial dimension, the theory for the spatial domain must be generalized to include higher-dimensional spatial Fourier transforms. However, this does not change the principles involved. Examples can be found in Sections 21.2 and 21.3 for the upward continuation of the Earth's gravity field.

15.8 Dereverberation filter

As an example of a filter that is derived in the time domain we consider here the description of reverberations in marine seismics. Suppose a seismic survey is carried out at sea. In such an experiment a ship tows a string of hydrophones that record the pressure variations just below the surface of the water, see Figure 15.3. Since the pressure vanishes at the surface of the water, the surface of the water totally reflects pressure waves and the reflection coefficient for reflection at the water surface is equal to -1. Let the reflection coefficient for waves reflecting upwards from the

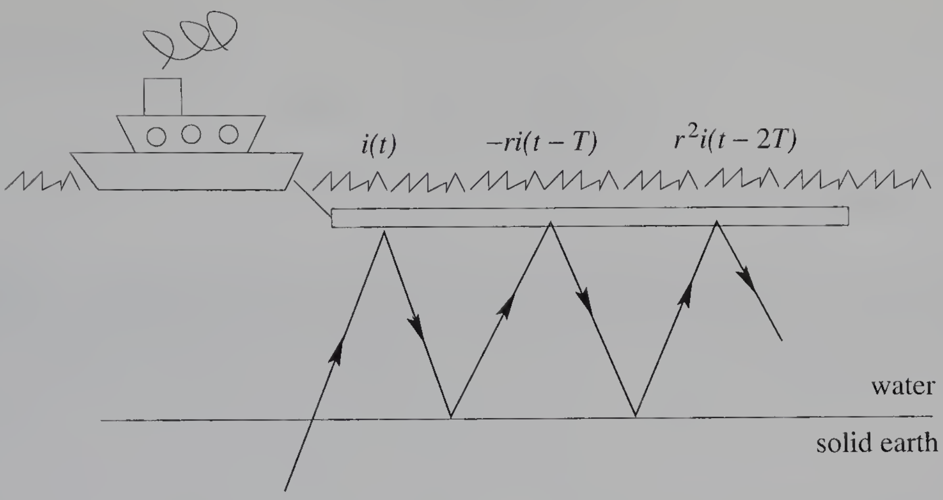

Fig. 15.3 The generation of reverberations in a marine seismic experiment.

sea bed be denoted by r. Since the contrast between the water and the solid earth below is not small, this reflection coefficient can be considerable. Because of energy conservation the reflected wave must be weaker than the incoming wave, so that $r \leq 1$.

Since the reflection coefficient of the sea bed is not small, waves can bounce back and forth repeatedly between the water surface and the sea bed. These reverberations are an unwanted artifact in seismic experiments. This is because a wave that has bounced back and forth in the water layer can be misinterpreted on a seismic section as a reflector in the Earth. For this reason one wants to eliminate these reverberations from seismic data.

Suppose the wavefield recorded by the hydrophones in the absence of reverberations is denoted by $i(t)$. Let the time it takes for a wave to travel from the water surface to the sea bed and back be denoted by T; this time is called the two-way travel time.

Problem a Show that a wave that has bounced back and forth once is given by $-ri(t-T)$. Hint: determine the amplitude of this wave from the reflection coefficients it encounters on its path and account for the time delay due to the bouncing up and down once in the water layer.

Problem b Generalize this result to a wave that bounces back and forth n-times in the water layer and show that the signal $o(t)$ recorded by the hydrophones is given by:

$$o(t) = i(t) - r\, i(t-T) + r^2\, i(t-2T) + \cdots$$

or

$$o(t) = \sum_{n=0}^{\infty} (-r)^n \, i(t - nT), \tag{15.59}$$

see Figure 15.3.

The notation $i(t)$ and $o(t)$ that was introduced in the previous section is deliberately used here. The action of the reverberation in the water layer is seen as a linear filter. The input of the filter $i(t)$ is the wavefield that would have been recorded if the waves did not bounce back and forth in the water layer. The output is the wavefield that results from the reverberations in the water layer. In a marine seismic experiment one records the wavefield $o(t)$ while one would like to know the signal $i(t)$ that contains just the reflections from below the water bottom. The process of removing the reverberations from the signal is called "dereverberation." The aim of this section is to derive a dereverberation filter that allows us to extract the input $i(t)$ from the recorded output $o(t)$.

Problem c It may not be obvious that (15.59) describes a linear filter of the form (15.57) that maps the input $i(t)$ onto the output $o(t)$. Show that (15.59) can be written in the form (15.57) with the impulse response $g(t)$ given by:

$$g(t) = \sum_{n=0}^{\infty} (-r)^n \, \delta(t - nT), \tag{15.60}$$

with $\delta(t)$ the Dirac delta function.

Problem d Show that $g(t)$ is indeed the impulse response. In other words, show that if a delta function is incident as a primary arrival at the water surface the reverberations within the water layer lead to the signal (15.60).

The deconvolution can most simply be carried out in the frequency domain. Let the Fourier transforms of $i(t)$ and $o(t)$, as defined by the transform (15.43), be denoted by $I(\omega)$ and $O(\omega)$ respectively. It follows from (15.59) that one needs to find the Fourier transform of $i(t - nT)$.

Problem e According to the definition (15.43) the Fourier transform of $i(t - \tau)$ is given by $(1/2\pi) \int_{-\infty}^{\infty} i(t - \tau) e^{i\omega t} dt$. Use a change of the integration variable to show that the Fourier transform of $i(t - \tau)$ is given by $I(\omega) e^{i\omega\tau}$.

What you have derived here is the *shift property* of the Fourier transform: a translation of a function over a time τ corresponds in the frequency domain to a multiplication by $e^{i\omega\tau}$

$$\left.\begin{array}{ccc} i(t) & \longleftrightarrow & I(\omega), \\ i(t-\tau) & \longleftrightarrow & I(\omega)e^{i\omega\tau}. \end{array}\right\} \tag{15.61}$$

Problem f Apply a Fourier transform to (15.59) for the output, use the shift property (15.61) for each term and show that the output in the frequency domain is related to the Fourier transform of the input by the following expression:

$$O(\omega) = \sum_{n=0}^{\infty} (-r)^n \, e^{i\omega nT} \, I(\omega). \tag{15.62}$$

Problem g Use the theory of Section 15.7 to show that the filter that describes the generation of reverberations is given in the frequency domain by:

$$G(\omega) = \frac{1}{2\pi} \sum_{n=0}^{\infty} (-r)^n \, e^{i\omega nT}. \tag{15.63}$$

Problem h Because of energy conservation, the reflection coefficient r is less than or equal to 1, therefore this series is guaranteed to converge. Sum this series to show that

$$G(\omega) = \frac{1}{2\pi} \frac{1}{1 + re^{i\omega T}}. \tag{15.64}$$

This is a useful result because it implies that the output and the input in the frequency domain are related by

$$O(\omega) = \frac{1}{1 + re^{i\omega T}} \, I(\omega). \tag{15.65}$$

Note that the action of the reverberation leads in the frequency domain to a simple division by $(1 + re^{i\omega T})$. Note also that (15.65) has a form similar to (3.40) which accounts for the reverberation of waves between two stacks of reflectors. This resemblance is no coincidence because the physics of waves bouncing back and forth between two reflectors is similar.

Problem i The goal of this section was to derive a dereverberation filter that produces $i(t)$ when $o(t)$ is given. Use (15.65) to derive a dereverberation filter in the frequency domain.

The dereverberation filter you have just derived is simple in the frequency domain; it only involves a multiplication of every frequency component $O(\omega)$ by a scalar.

Since multiplication is a simple and efficient procedure it is attractive to carry out dereverberation in the frequency domain. The dereverberation filter you have just derived was developed originally by Backus [7].

The simplicity of the dereverberation filter hides a nasty complication. If the reflection coefficient r and the two-way travel time T are known exactly and if the sea bed is exactly horizontal there is no problem with the dereverberation filter. However, in practice one only has *estimates* of these quantities. Let these estimates be denoted by r' and T' respectively. The reverberations lead in the frequency domain to a division by $1 + re^{i\omega T}$, while the dereverberation filter based on the estimated parameters leads to a multiplication with $1 + r'e^{i\omega T'}$. The net effect of the generation of the reverberations and the subsequent dereverberation is thus given in the frequency domain by a multiplication by

$$\frac{1 + r'e^{i\omega T'}}{1 + re^{i\omega T}} .$$

Problem j Show that when the reflection coefficients are close to unity and when the estimate of the travel time is not accurate ($T' \neq T$) the term given above differs appreciably from unity. Explain why this implies that the dereverberation does not work well.

In practice one faces not only the problem that the estimates of the reflection coefficients and the two-way travel time may be inaccurate, but in addition the sea bed may not be exactly flat and there may be variations in the reflection coefficient along the sea bed. In which case the performance of the dereverberation filter can be significantly degraded.

15.9 Design of frequency filters

In this section we consider the problem in which a time series $i(t)$ is recorded and is contaminated with high-frequency noise. The aim of this section is to derive a filter in the time domain that removes the frequency components with a frequency greater than a cut-off frequency ω_0 from the time series. Such a filter is called a low-pass filter because only frequency components lower than the threshold ω_0 pass the filter.

Problem a Show that this filter is given in the frequency domain by:

$$G(\omega) = \begin{cases} 1 & \text{if} \quad |\omega| \leq \omega_0 \\ 0 & \text{if} \quad |\omega| > \omega_0 \end{cases} . \tag{15.66}$$

Problem b Explain why the absolute value of the frequency should be used in this expression.

Problem c Show that this filter is given in the time domain by

$$g(t) = \int_{-\omega_0}^{\omega_0} e^{-i\omega t} \, d\omega. \tag{15.67}$$

Problem d Carry out the integration over frequency to derive that the filter is explicitly given by

$$g(t) = 2\omega_0 \operatorname{sinc}(\omega_0 t), \tag{15.68}$$

where the sinc function is defined by

$$\operatorname{sinc} x \equiv \frac{\sin x}{x}. \tag{15.69}$$

Problem e Sketch the impulse response (15.68) of the low-pass filter as a function of time. Determine the behavior of the filter for $t = 0$ and show that the first zero-crossing of the filter is at time $t = \pm\pi/\omega_0$.

The zero-crossing of the filter is of fundamental importance. It implies that the width of the impulse response in the time domain is given by $2\pi/\omega_0$.

Problem f Use expression (15.66) to show that the width of the filter in the frequency domain is given by $2\omega_0$.

This means that when the the cut-off frequency ω_0 is increased, the width of the filter in the frequency domain increases but the width of the filter in the time domain decreases. A large width of the filter in the frequency domain corresponds to a small width of the filter in the time domain and vice versa.

Problem g Show that the product of the width of the filter in the time domain and the width of the same filter in the frequency domain is given by 4π.

The significance of this result is that this product is independent of the frequency ω_0. This implies that the filter cannot be arbitrarily peaked in both the time domain and the frequency domain. This effect has pronounced consequences since it is the essence of the *uncertainty relation of Heisenberg* which states that the position and momentum of a particle can never be known exactly; more details can be found in the book of Merzbacher [69].

The filter (15.68) does not actually have very desirable properties; it has two basic problems. The first problem is that the filter decays slowly with time. This means that the length of the filter in the time domain is very great, and hence the convolution of a time series with the filter is numerically a rather inefficient process. This can be solved by making the cut-off of the filter in the frequency domain more gradual than the frequency cut-off defined in (15.66), for example by using the filter $G(\omega) = (1 + |\omega| / \omega_0)^{-n}$ with n a positive integer.

Problem h Does this filter have the steepest cut-off for low values of n or for high values of n? Hint: make a plot of $G(\omega)$ as a function of ω.

The second problem is that the filter is not *causal*. This means that when a function is convolved with the filter (15.68), the output of the filter depends on the value of the input at later times, that is the filter output depends on the input in the future.

Problem i Show that this is the case, and that the output depends on the input of earlier times only when $g(t) = 0$ for $t < 0$.

A causal filter can be designed by using the theory of analytic functions described in Chapter 16. The design of filters is quite an art; details can be found for example in the books of Robinson and Treitel [90] or Claerbout [27].

15.10 Linear filters and linear algebra

There is a close analogy between the theory of linear filters of Section 15.7 and the eigenvector decomposition of a matrix in linear algebra as treated in Section 13.4. To see this we will use the same notation as in Section 15.7 and use the Fourier transform (15.45) to write the output of the filter in the time domain as:

$$o(t) = \int_{-\infty}^{\infty} O(\omega) e^{-i\omega t} d\omega. \tag{15.70}$$

Problem a Use expression (15.58) to show that this can be written as

$$o(t) = 2\pi \int_{-\infty}^{\infty} G(\omega) I(\omega) e^{-i\omega t} d\omega, \tag{15.71}$$

and carry out an inverse Fourier transformation of $I(\omega)$ to derive that

$$o(t) = \iint_{-\infty}^{\infty} G(\omega)e^{-i\omega t}e^{i\omega\tau}i(\tau)\,d\omega d\tau. \tag{15.72}$$

In order to establish the connection with linear algebra we introduce by analogy with (15.33) the following basis functions:

$$u_\omega(t) \equiv \frac{1}{\sqrt{2\pi}}e^{-i\omega t}, \tag{15.73}$$

and the inner product

$$(f \cdot g) \equiv \int_{-\infty}^{\infty} f^*(t)g(t)\,dt. \tag{15.74}$$

Problem b Use the results of Section 15.4 to show that these basis functions are orthonormal for this inner product in the sense that

$$\left(u_\omega \cdot u_{\omega'}\right) = \delta(\omega - \omega'). \tag{15.75}$$

Problem c These functions play the same role as the eigenvectors in Section 13.4. To which expression in Section 13.4 does the above expression correspond?

Problem d Show that (15.72) can be written as

$$o(t) = \int_{-\infty}^{\infty} G(\omega)u_\omega(t)\,(u_\omega \cdot i)\,d\omega. \tag{15.76}$$

This expression should be compared with

$$\mathbf{A}\mathbf{p} = \sum_{n=1}^{N} \lambda_n \hat{\mathbf{v}}^{(n)}\left(\hat{\mathbf{v}}^{(n)} \cdot \mathbf{p}\right). \tag{13.44}$$

The integration over frequency plays the same role as the summation over eigenvectors in (13.44). Expression (15.76) can be seen as a description for the operator $g(t)$ in the time domain that maps the input function $i(t)$ onto the output $o(t)$.

Problem e Use (15.57), (15.74), and (15.76) to show that:

$$g(t-\tau) = 2\pi \int_{-\infty}^{\infty} G(\omega)u_\omega(t)u_\omega^*(\tau)\,d\omega. \tag{15.77}$$

There is a close analogy between this expression and the dyadic decomposition of a matrix into its eigenvectors and eigenvalues derived in Section 13.4.

Problem f To see this connection show that (13.45) can be written in component form as:

$$A_{ij} = \sum_{n=1}^{N} \lambda_n \hat{v}_i^{(n)} \hat{v}_j^{(n)T} . \tag{15.78}$$

The sum over eigenvalues in (15.78) corresponds to the integration over frequency in (15.77). In Section 13.4 linear algebra in a finite-dimensional vector space was treated. In such a space there is a finite number of eigenvalues. In this section, a function space with infinitely many degrees of freedom is analyzed; it will come as no surprise that for this reason the sum over a finite number of eigenvalues should be replaced by an integration over the continuous variable ω. The index i in (15.78) corresponds to the variable t in (15.77) while the index j corresponds to the variable τ.

Problem g Establish the connection between *all* variables in (15.77) and (15.78). Show specifically that $G(\omega)$ plays the role of eigenvalue λ_n and u_ω plays the role of eigenvector. Which operation in (15.77) corresponds to the transpose that is taken of the second eigenvector in (15.78)?

You may wonder why the function $u_\omega(t) = e^{-i\omega t}/\sqrt{2\pi}$, defined in (15.73), and not some other function plays the role of the eigenvector of the impulse response operator $g(t - \tau)$. To see why this is so we have to understand what a linear filter actually does. Let us first consider the example of the reverberation filter of Section 15.8. According to (15.59) the reverberation filter is given by:

$$o(t) = i(t) - ri(t - T) + r^2 i(t - 2T) + \cdots . \tag{15.59}$$

It follows from this expression that what the filter really does is to take the input $i(t)$, translate it over a time nT to a new function $i(t - nT)$, multiply each term with $(-r)^n$ and sum over all values of n. This means that the filter is a combination of three operations: (i) translation in time, (ii) multiplication, and (iii) summation over n. The same conclusion holds for any general time-invariant linear filter.

Problem h Use a change of the integration variable to show that the action of a time-invariant linear filter as given in (15.57) can be written as

$$o(t) = \int_{-\infty}^{\infty} g(\tau) i(t-\tau)\, d\tau. \tag{15.79}$$

The function $i(t-\tau)$ is the function $i(t)$ translated over a time τ. This translated function is multiplied by $g(\tau)$ and an integration over all values of τ is carried out. This means that in general the action of a linear filter can be seen as a combination of translation in time, multiplication, and integration over all translations τ. How can this be used to explain that the correct eigenfunctions to be used are $u_\omega(t) = e^{-i\omega t}/\sqrt{2\pi}$? The answer does not lie in the multiplication because *any* function is an eigenfunction of the operator that carries out multplication by a constant, that is $af(t) = \lambda f(t)$ for every function $f(t)$ with the eigenvalue $\lambda = a$. This means that the translation operator is the reason that the eigenfunctions are $u_\omega(t) = e^{-i\omega t}/\sqrt{2\pi}$. Let the operator that carries out a translation over a time τ be denoted by T_τ:

$$T_\tau f(t) \equiv f(t-\tau). \tag{15.80}$$

Problem i Show that the functions $u_\omega(t)$ defined in (15.73) are the eigenfunctions of the translation operator T_τ, that is show that $T_\tau u_\omega(t) = \lambda u_\omega(t)$. Express the eigenvalue λ of the translation operator in terms of the translation time τ.

Problem j Compare this result with the shift property of the Fourier transform that was derived in (15.61).

This means that the functions $u_\omega(t)$ are the eigenfunctions to be used for the eigenfunction decomposition of a linear time-invariant filter, because these functions are eigenfunctions of the translation operator.

Problem k You identified in problem e the eigenvalues of the filter with $G(\omega)$. Show that this interpretation is correct: in other words show that when the filter g acts on the function $u_\omega(t)$ the result can be written as $G(\omega)u_\omega(t)$. Hint: go back to problem e of Section 15.7.

This analysis shows that the Fourier transform, which uses the functions $e^{-i\omega t}$, is so useful because these functions are the eigenfunctions of the translation operator. However, this also points to a limitation of the Fourier transform. Consider a linear filter that is not time-invariant, that is a filter in which the output does not depend only on the *difference* between the input time τ and the output time t. Such a filter satisfies the general equation (15.55) rather than the convolution integral (15.57). The action of a filter that is not time-invariant *cannot* in general be written as a combination of the following operations: multiplication, translation, and integration. This means that for such a filter the functions $e^{-i\omega t}$ that form the basis of the Fourier transform are not the appropriate eigenfunctions. The upshot of this is that in practice the Fourier transform is only useful for systems that are time-invariant, or in general that are translationally invariant in the coordinate that is used.

16

Analytic functions

In this chapter we consider complex functions in the complex plane. The reason for doing this is that the requirement that the function "behaves well" (this is defined later) imposes remarkable constraints on such complex functions. Since these constraints coincide with some of the laws of physics, the theory of complex functions has a number of important applications in mathematical physics. In this chapter complex functions $h(z)$ are treated that are decomposed into real and imaginary parts:

$$h(z) = f(z) + ig(z); \tag{16.1}$$

hence the functions $f(z)$ and $g(z)$ are assumed to be real. The complex number z is written as $z = x + iy$, so that $x = \Re e(z)$ and $y = \Im m(z)$, where $\Re e$ and $\Im m$ denote the real and imaginary part respectively.

16.1 Theorem of Cauchy–Riemann

Let us first consider a real function $F(x)$ of a real variable x. The derivative of such a function is defined by the rule

$$\frac{dF}{dx} = \lim_{\Delta x \to 0} \frac{F(x + \Delta x) - F(x)}{\Delta x}. \tag{16.2}$$

In general there are two ways in which Δx can approach zero: from above and from below. For a function that is differentiable it does not matter whether Δx approaches zero from above or from below. If the limits $\Delta x \downarrow 0$ and $\Delta x \uparrow 0$ do give a different result it is a sign that the function does not behave well, it has a kink and the derivative is not unambiguously defined, see Figure 16.1.

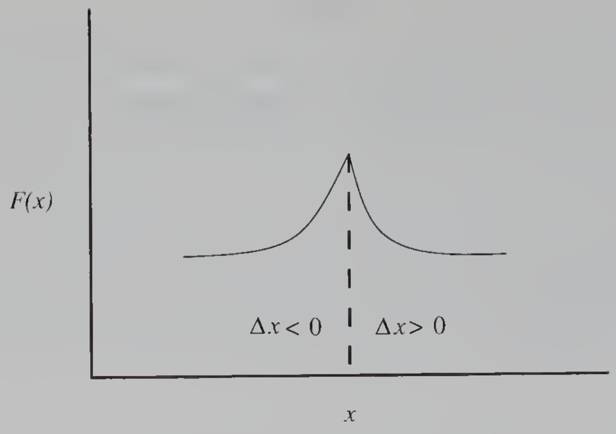

Fig. 16.1 A function $F(x)$ that is not differentiable.

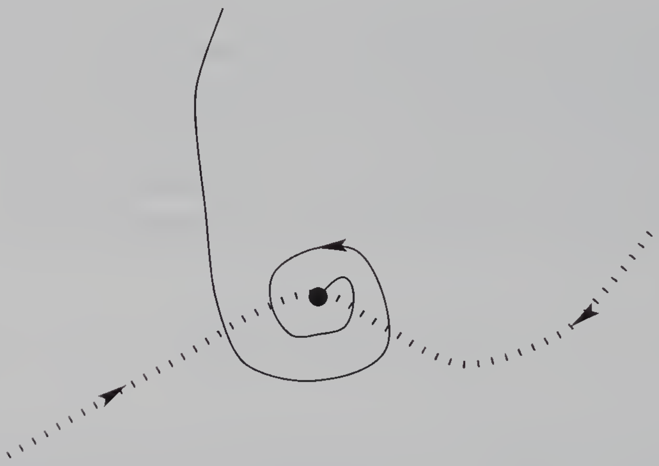

Fig. 16.2 Examples of paths along which the limit $\Delta z \to 0$ can be taken.

For complex functions the derivative is defined in the same way as in (16.1) for real functions:

$$\frac{dh}{dz} = \lim_{\Delta z \to 0} \frac{h(z + \Delta z) - h(z)}{\Delta z}. \tag{16.3}$$

For real functions, Δx could approach zero in two ways: from below and from above. However, the limit $\Delta z \to 0$ in (16.3) can be taken in infinitely many ways. As an example see Figure 16.2 where some different paths that one can use to let Δz approach zero are sketched. These do not necessarily give the same result.

Problem a Consider the function $h(z) = e^{1/z}$. Using the definition (16.3) compute dh/dz at the point $z = 0$ when Δz approaches zero: (i) from the positive real axis, (ii) from the negative real axis, (iii) from the positive imaginary axis, and (iv) from the negative imaginary axis.

You have discovered that for some functions the result of the limit Δz depends critically on the path that one uses in the limit process. The derivative of such a function is not defined unambiguously. However, for many functions the value of the derivative does *not* depend on the way that Δz approaches zero. When these functions and their derivatives are also finite, they are called *analytic functions.* The requirement that the derivative does not depend on the way in which Δz approaches zero imposes a constraint on the real and imaginary parts of the complex function. To see this we will let Δz approach zero along the real axis and along the imaginary axis.

Problem b Consider a complex function of the form (16.1) and compute the derivative dh/dz by setting $\Delta z = \Delta x$ with Δx a real number. (Hence Δz approaches zero along the real axis.) Show that the derivative is given by $dh/dz = \partial f/\partial x + i\partial g/\partial x$.

Problem c Compute the derivative dh/dz also by setting $\Delta z = i\Delta y$ with Δy a real number. (Hence Δz approaches zero along the imaginary axis.) Show that the derivative is given by $dh/dz = \partial g/\partial y - i\partial f/\partial y$.

Problem d When $h(z)$ is analytic these two expressions for the derivative are by definition equal. Show that this implies that:

$$\frac{\partial f}{\partial x} = \frac{\partial g}{\partial y}, \tag{16.4}$$

$$\frac{\partial g}{\partial x} = -\frac{\partial f}{\partial y}. \tag{16.5}$$

These are puzzling expressions since (16.4) and (16.5) imply that the real and imaginary parts of an analytic complex function are not independent of each other: they are coupled by the constraints imposed by equations (16.4) and (16.5). Expressions (16.4) and (16.5) are called the *Cauchy–Riemann relations.*

Problem e Use these relations to show that both $f(x, y)$ and $g(x, y)$ are harmonic functions. These are functions for which the Laplacian vanishes:

$$\nabla^2 f = \nabla^2 g = 0. \tag{16.6}$$

Hence we have found not only that f and g are coupled to each other; in addition the functions f and g must be harmonic functions. This is exactly the reason why this theory is so useful in mathematical physics because harmonic functions arise in several applications, see the examples in the coming sections. However, we have not found all the properties of harmonic functions yet.

Problem f Show that:

$$(\nabla f \cdot \nabla g) = 0. \tag{16.7}$$

Since the gradient of a function is perpendicular to the lines where the function is constant this implies that the curves where f is constant and where g is constant intersect each other at a fixed angle.

Problem g Determine this angle.

Problem h Verify the properties (16.4)–(16.7) explicitly for the function $h(z) = z^2$. Also sketch the lines in the complex plane where $f = \Re e(h)$ and $g = \Im m(h)$ are constant.

We have still not fully explored all the properties of analytic functions. Let us consider a line integral $\oint_C h(z)dz$ along a closed contour C in the complex plane.

Problem i Use the property $dz = dx + idy$ to deduce that:

$$\oint_C h(z)dz = \oint_C \mathbf{v} \cdot d\mathbf{r} + i \oint_C \mathbf{w} \cdot d\mathbf{r}, \tag{16.8}$$

where $d\mathbf{r} = \begin{pmatrix} dx \\ dy \end{pmatrix}$ and the vectors $\mathbf{v}$ and $\mathbf{w}$ are defined by:

$$\mathbf{v} = \begin{pmatrix} f \\ -g \end{pmatrix}, \quad \mathbf{w} = \begin{pmatrix} g \\ f \end{pmatrix}. \tag{16.9}$$

Note that we are now using x and y in a dual role as the real and imaginary parts of a complex number, as well as the Cartesian coordinates in a plane. In the following problem we will (perhaps confusingly) use the notation z both for a complex number in the complex plane, as well as for the familiar z-coordinate in a three-dimensional Cartesian coordinate system.

Problem j Show that the Cauchy–Riemann relations (16.4)–(16.5) imply that the z-component of the *curl* of $\mathbf{v}$ and $\mathbf{w}$ vanishes: $(\nabla \times \mathbf{v})_z = (\nabla \times \mathbf{w})_z = 0$, and use (16.8) and the theorem of Stokes (9.2) to show that when $h(z)$ is analytic everywhere within the contour C:

$$\oint_C h(z)\,dz = 0, \qquad \text{where } h(z) \text{ is analytic within } C. \qquad (16.10)$$

This means that the line integral of a complex functions along *any* contour which encloses a region of the complex plane where that function is analytic is equal to zero. We will make extensive use of this property in Chapter 17 where we treat integration in the complex plane.

16.2 Electric potential

Analytic functions are often useful in the determination of the electric field and the potential for two-dimensional problems. The electric field satisfies the field equation (8.13): $(\nabla \cdot \mathbf{E}) = \rho(\mathbf{r})/\varepsilon_0$. In free space the charge-density vanishes, hence $(\nabla \cdot \mathbf{E}) = 0$. The electric field is related to the potential V through the relation

$$\mathbf{E} = -\nabla V. \qquad (16.11)$$

Problem a Show that in free space the potential is a harmonic function:

$$\nabla^2 V(x, y) = 0. \qquad (16.12)$$

We can exploit the theory of analytic functions by noting that the real and imaginary parts of analytic functions both satisfy (16.12). This implies that if we take $V(x, y)$ to be the real part of a complex analytic function $h(x + iy)$, (16.12) is automatically satisfied.

Problem b It follows from (16.11) that the electric field is perpendicular to the lines where V is constant. Show that this implies that the electric field lines are also perpendicular to the lines $V = const$. Use the theory of the previous section to argue that the field lines are the lines where the imaginary part of $h(x + iy)$ is constant.

This means that we receive a bonus by expressing the potential V as the real part of a complex analytic function, because the field lines simply follow from the requirement that $\Im m(h) = const.$

Suppose we want to know the potential in the half-space $y \geq 0$ when we have specified the potential on the x-axis. (Mathematically this means that we want to solve the equation $\nabla^2 V = 0$ for $y \geq 0$ when $V(x, y = 0)$ is given.) If we can find an analytic function $h(x + iy)$ such that on the x-axis (where $y = 0$) the real part of h is equal to the potential, we have solved our problem because the real part of h by definition satisfies the required boundary condition and it satisfies the field equation (16.12).

Problem c Consider a potential that is given on the x-axis by

$$V(x, y=0) = V_0 \, e^{-x^2/a^2} . \tag{16.13}$$

Show that on the x-axis this function can be written as $V = \Re e(h)$ with

$$h(z) = V_0 \, e^{-z^2/a^2} . \tag{16.14}$$

Problem d This means that we can determine the potential and the field lines throughout the half-plane $y \geq 0$. Use the theory of this section to show that the potential is given by

$$V(x, y) = V_0 \, e^{(y^2-x^2)/a^2} \cos\left(\frac{2xy}{a^2}\right) . \tag{16.15}$$

Problem e Verify explicitly that this solution satisfies the boundary condition at the x-axis and that it satisfies the field equation (16.12).

Problem f Show that the field lines are given by the relation

$$e^{(y^2-x^2)/a^2} \sin\left(\frac{2xy}{a^2}\right) = const. \tag{16.16}$$

Problem g Sketch the field lines and the lines where the potential is constant in the half-space $y \geq 0$.

In this derivation we have extended the solution $V(x, y)$ into the upper half-plane by identifying it with an analytic function in the half-plane which has on the x-axis a real part which equals the potential on the x-axis. Note that we found the solution to this problem without explicitly solving the partial differential equation (16.12) which governs the potential. The approach we have taken is called *analytic continuation* since we continue an analytic function from one region (the x-axis) into the upper half-plane. Analytic continuation turns out to be a very unstable process. This can be verified explicitly for this example.

Problem h Sketch the potential $V(x, y)$ as a function of x for the values $y = 0$, $y = a$, and $y = 10a$. What is the wavelength of the oscillations in the x-direction of the potential $V(x, y)$ for these values of y? Show that when we slightly perturb the constant a the perturbation of the potential increases for increasing values of y. This implies that when we slightly perturb the boundary condition the solution is more perturbed as we move further away from that boundary.

16.3 Fluid flow and analytic functions

As a second application of the theory of analytic functions we consider fluid flow. At the end of Section 6.2 we saw that the streamlines of the flow can be determined by solving the differential equations $d\mathbf{r}/dt = \mathbf{v}(\mathbf{r})$ for the velocity field (6.16)–(6.17). This requires the solution of a system of nonlinear differential equations which is very difficult. Here the theory of analytic functions is used to solve this problem in a simple way. We consider once again a fluid that is incompressible ($(\nabla \cdot \mathbf{v}) = 0$) and specialize to the special case in which the vorticity of the flow vanishes:

$$\nabla \times \mathbf{v} = 0. \tag{16.17}$$

Such a flow is called *irrotational* because it does not rotate (see Sections 7.3 and 7.4). The requirement (16.17) is automatically satisfied when the velocity is the gradient of a scalar function f:

$$\mathbf{v} = \nabla f. \tag{16.18}$$

Problem a Show this by taking the curl of the previous expression.

The function f plays the same role for the velocity field $\mathbf{v}$ as the electric potential V does for the electric field $\mathbf{E}$. For this reason, flow with a vorticity equal to zero is called *potential flow*.

Problem b Show that the requirement that the flow is incompressible implies that

$$\nabla^2 f = 0. \tag{16.19}$$

We now specialize to the special case of incompressible and irrotational flow in two dimensions. In that case we can again use the theory of analytic functions to describe the flow field. Once we can identify the potential $f(x, y)$ with the real part of an analytic function $h(x + iy)$ we know that (16.19) must be satisfied.

Problem c Consider the velocity field (6.13) due to a point source at $\mathbf{r} = 0$. Show that this flow field follows from the potential

$$f(x, y) = \frac{\dot{V}}{2\pi} \ln r, \tag{16.20}$$

where $r = \sqrt{x^2 + y^2}$.

Problem d Verify explicitly that this potential satisfies (16.19) except for $r = 0$. For what physical reason is (16.19) not satisfied at $r = 0$?

We now want to identify the potential $f(x, y)$ with the real part of an analytic function $h(x + iy)$. We know already that the flow follows from the requirement that the velocity is the gradient of the potential, hence it follows by taking the gradient of the real part of h. The curves $f = const.$ are perpendicular to the flow because $\mathbf{v} = \nabla f$ is perpendicular to these curves. However, it was shown in Section 16.1 that the curves $g = \Im m(h) = const.$ are perpendicular to the curves $f = \Re e(h) = const.$ This means that the flow lines are given by the curves $g = \Im m(h) = const.$ In order to use this, we first have to find an analytic function with a real part given by (16.20). The simplest guess is to replace r in (16.20) by the complex variable z:

$$h(z) = \frac{\dot{V}}{2\pi} \ln z. \tag{16.21}$$

Problem e Verify that for complex z the real part of this function indeed satisfies (16.20) and that the imaginary part g of this function is given by:

$$g = \frac{\dot{V}}{2\pi} \varphi, \tag{16.22}$$

where $\varphi = \arctan(y/x)$ is the argument of the complex number. Hint: use the representation $z = re^{i\varphi}$ for the complex numbers.

Problem f Sketch the lines $g(x, y) = const.$ and verify that these lines indeed represent the flow lines in this flow.

Now we consider the more complicated problem in Section 6.2 in which the flow has a source at $\mathbf{r}_+ = (L, 0)$ and a sink at $\mathbf{r}_- = (-L, 0)$. The velocity field is given by (6.16) and (6.17) and our goal is to determine the flow lines without solving the differential equation $d\mathbf{r}/dt = \mathbf{v}(\mathbf{r})$. The points $\mathbf{r}_+$ and $\mathbf{r}_-$ can be represented in the complex plane by the complex numbers $z_+ = L + i0$ and $z_- = -L + i0$ respectively. For the source at $\mathbf{r}_+$ flow is represented by the potential $(\dot{V}/2\pi) \ln |z - z_+|$; this follows from the solution (16.21) for a single source by moving this source to $z = z_+$.

Problem g Using a similar reasoning determine the contribution to the potential f by the sink at $\mathbf{r}_-$. Show that the potential of the total flow is given by:

$$f(z) = \frac{\dot{V}}{2\pi} \ln \left(\frac{|z - z_+|}{|z - z_-|} \right). \tag{16.23}$$

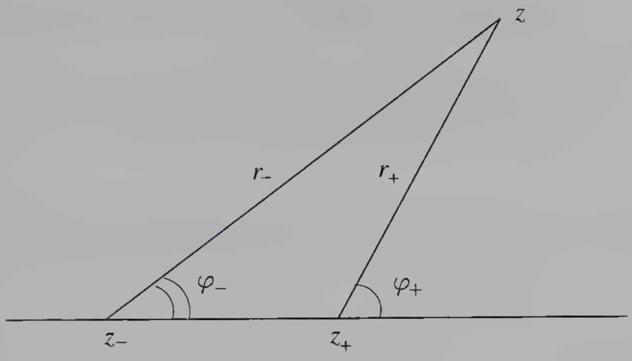

Fig. 16.3 Definition of the geometric variables for the fluid flow with a source and a sink.

Problem h Express this potential in x and y, compute the gradient, and verify that this potential indeed gives the flow of (6.16)–(6.17). You may find Figure 16.3 helpful.

We have found that the potential f is the real part of the complex function

$$h(z) = \frac{\dot{V}}{2\pi} \ln \left(\frac{z - z_+}{z - z_-} \right). \tag{16.24}$$

Problem i Write $z - z_\pm = r_\pm e^{i\varphi_\pm}$ (with $r_\pm$ and $\varphi_\pm$ defined in Figure 16.3), take the imaginary part of (16.24), and show that $g = \Im m(h)$ is given by:

$$g = \frac{\dot{V}}{2\pi} \left(\varphi_+ - \varphi_- \right). \tag{16.25}$$

Problem j Show that the streamlines of the flow are given by the relation

$$\arctan \left(\frac{y}{x - L} \right) - \arctan \left(\frac{y}{x + L} \right) = const. \tag{16.26}$$

A plot of the streamlines can thus be drawn by making a contour plot of the function on the left-hand side of (16.26). This treatment is simpler than solving the differential equation $d\mathbf{r}/dt = \mathbf{v}(\mathbf{r})$.

17

Complex integration

In Chapter 16 the properties of analytic functions in the complex plane were treated. One of the key results is that the contour integral of a complex function is equal to zero when the function is analytic everywhere in the area of the complex plane enclosed by that contour, see (16.10). From this it follows that the integral of a complex function along a closed contour is nonzero only when the function is not analytic in the area enclosed by the contour. Functions that are not analytic come in different types. In this section complex functions are considered that are not analytic only at isolated points. These points where the function is not analytic are called the *poles* of the function.

17.1 Nonanalytic functions

When a complex function is analytic at a point z_0, it can be expanded in a Taylor series around that point. This implies that within a certain region around z_0 the function can be written as:

$$h(z) = \sum_{n=0}^{\infty} a_n (z - z_0)^n . \tag{17.1}$$

Note that in this sum only positive powers of $(z - z_0)$ appear.

Problem a Show that the function $h(z) = \sin(z)/z$ can be written as a Taylor series around the point $z_0 = 0$ of the form (17.1) and determine the coefficients a_n.

Not all functions can be represented in a series of the form (17.1). As an example consider the function $h(z) = e^{1/z}$. The function is not analytic at the point $z = 0$

(why?). Expanding the exponential leads to the expansion

$$h(z) = e^{1/z} = 1 + \frac{1}{z} + \frac{1}{2!}\frac{1}{z^2} + \cdots = \sum_{n=0}^{\infty} \frac{1}{n!}\frac{1}{z^n}. \tag{17.2}$$

In this case an expansion in *negative* powers of z is needed to represent this function. Of course, each of the terms $1/z^n$ is for $n \geq 1$ singular at $z = 0$; this reflects the fact that the function $e^{1/z}$ has a pole at $z = 0$. In this section we consider complex functions which can be expanded around a point z_0 as an infinite sum of *integer* powers of $(z - z_0)$:

$$h(z) = \sum_{n=-\infty}^{\infty} a_n (z - z_0)^n. \tag{17.3}$$

It should be noted, however, that not every function can be represented as such a sum. An example is the function $h(z) = \sqrt{z}$. This function can be written as $h(z) = z^{1/2}$. When viewed as a sum of powers of z^n, the term $z^{1/2}$ is the only term that contributes. Since $n = \frac{1}{2}$ is not an integer, this function cannot be written in the form (17.3).

17.2 Residue theorem

It was argued at the beginning of this chapter that the integral of a complex function around a closed contour in the complex plane is only nonzero when the function is not analytic at some point in the area enclosed by the contour. In this section we derive the value of the contour integral. Let us integrate a complex function $h(z)$ along a contour C in the complex plane that encloses a pole of the function at the point z_0, see the left-hand panel of Figure 17.1. Note that the integration is carried out in the counterclockwise direction. It is assumed that around the point z_0 the

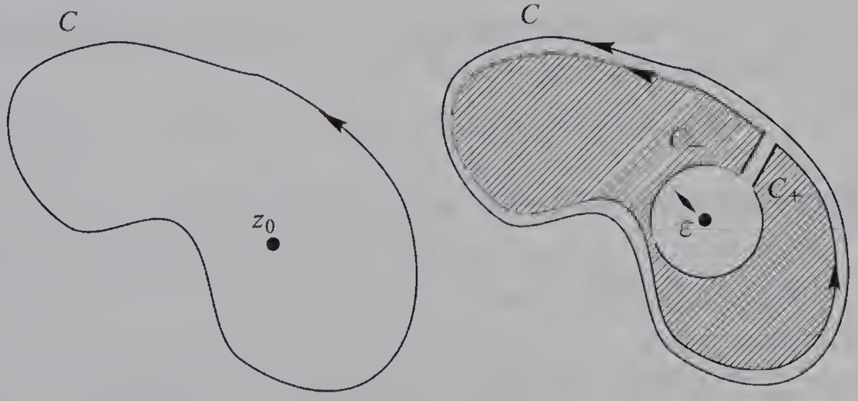

Fig. 17.1 Definition of the contours for the contour integration.

function $h(z)$ can be expanded in a power series of the form (17.3). It is our goal to determine the value of the contour integral $\oint_C h(z)dz$.

The first step in the determination of the contour integral is to recognize that within the shaded area in the right-hand panel of Figure 17.1 the function $h(z)$ is analytic because we have assumed that $h(z)$ is only nonanalytic at the point z_0. By virtue of the identity (16.10) this implies that

$$\oint_{C^*} h(z)\,dz = 0, \tag{17.4}$$

where the path C^* consists of the contour C, a small circle with radius ε around z_0 and the paths C^+ and C^- in the right-hand panel of Figure 17.1.

Problem a Show that the integrals along C^+ and C^- do not give a net contribution to the total integral:

$$\int_{C^+} h(z)\,dz + \int_{C^-} h(z)\,dz = 0. \tag{17.5}$$

Hint: note the opposite directions of integration along the paths C^+ and C^-.

Problem b Use this result and (17.4) to show that the integral along the original contour C is identical to the integral along the small circle C_ε around the point where $h(z)$ is not analytic:

$$\oint_C h(z)\,dz = \oint_{C_\varepsilon} h(z)\,dz. \tag{17.6}$$

Problem c The integration along C is in the counterclockwise direction. Is the integration along C_ε in the clockwise or in the counter-clockwise direction?

Expression (17.6) is useful because the integral along the small circle can be evaluated by using that close to z_0 the function $h(z)$ can be written as the series (17.3). When one does this the integration path C_ε needs to be parameterized. This can be achieved by writing the points on the path C_ε as

$$z = z_0 + \varepsilon e^{i\varphi}\ , \tag{17.7}$$

with φ running from 0 to 2π since C_ε is a complete circle.

Problem d Use (17.3), (17.6), and (17.7) to derive that

$$\oint_C h(z)\,dz = \sum_{n=-\infty}^{\infty} i a_n \varepsilon^{(n+1)} \int_0^{2\pi} e^{i(n+1)\varphi}\,d\varphi. \tag{17.8}$$

This is useful because it expresses the contour integral in the coefficients a_n of the expansion (17.3). It turns out that only the coefficient a_{-1} gives a nonzero contribution.

Problem e Show by direct integration that:

$$\int_0^{2\pi} e^{im\varphi} d\varphi = \begin{cases} 0 & \text{for} \quad m \neq 0 \\ 2\pi & \text{for} \quad m = 0 \end{cases} . \tag{17.9}$$

Problem f Use this result to derive that only the term $n = -1$ contributes to the sum on the right-hand side of (17.8) and that

$$\oint_C h(z)\,dz = 2\pi i a_{-1} \tag{17.10}$$

It may seem surprising that only the term $n = -1$ contributes to the sum on the right-hand side of (17.8). However, we could have anticipated this result because we had already discovered that the contour integral does not depend on the precise choice of the integration path. It can be seen that in the sum (17.8) each term is proportional to $\varepsilon^{(n+1)}$. Since we know that the integral does not depend on the choice of the integration path, and hence on the size of the circle C_ε, one would expect that only terms which do not depend on the radius ε contribute. This is only the case when $n + 1 = 0$, hence only for the term $n = -1$ is the contribution independent of the size of the circle. It is indeed only this term that gives a nonzero contribution to the contour integral.

The coefficient a_{-1} is usually called the *residue* and is denoted by the symbol Res $h(z_0)$ rather than a_{-1}. However, remember that there is nothing mysterious about the residue, it is simply defined as

$$\text{Res}\; h(z_0) \equiv a_{-1}. \tag{17.11}$$

With this definition the result (17.10) can trivially be written as

$$\oint_C h(z)\,dz = 2\pi i \text{Res}\; h(z_0) \qquad \text{(counterclockwise direction).} \tag{17.12}$$

This may appear to be a rather uninformative rewrite of (17.10) but it is the form (17.12) that you will find in the literature. The identity (17.12) is called the *residue theorem*.

Of course the residue theorem is only useful when one can determine the coefficient a_{-1} in the expansion (17.3). You can find in Section 2.12 of Butkov [24] an overview of methods for computing the residue. Here we present the two most widely used ones. The first method is to determine the power series expansion (17.3) of the function explicitly.

Problem g Determine the power series expansion of the function $h(z) = \sin(z)/z^4$ around $z = 0$ and use this expansion to find the residue.

Unfortunately, this method does not always work. For some special functions other tricks can be used. Here we consider functions with a simple pole; these are functions where the terms for $n < -1$ do not contribute in the expansion (17.3):

$$h(z) = \sum_{n=-1}^{\infty} a_n (z - z_0)^n . \tag{17.13}$$

An example of such a function is $h(z) = \cos(z)/z$. The residue at the point z_0 follows by "extracting" the coefficient a_{-1} from (17.13).

Problem h Multiply (17.13) by $(z - z_0)$, and take the limit $z \to z_0$ to show that:

$$\text{Res}\, h(z_0) = \lim_{z \to z_0} (z - z_0) h(z) \qquad \text{(simple pole).} \tag{17.14}$$

However, remember that this recipe works only for functions with a simple pole, it gives the wrong answer (infinity) when applied to a function which has nonzero coefficients a_n for $n < -1$ in (17.3).

In the treatment in this section we have considered an integration in the counterclockwise direction around the pole z_0.

Problem i Repeat the derivation of this section for a contour integration in the *clockwise* direction around the pole z_0 and show that in that case

$$\oint_C h(z)\, dz = -2\pi i \text{Res}\, h(z_0) \qquad \text{(clockwise direction).} \tag{17.15}$$

Find out in which step of the derivation the minus sign is picked up!

Problem j It may happen that a contour encloses not a single pole but a number of poles at points z_j. Find for this case a contour similar to the contour C^* in the right-hand panel of Figure 17.1 to show that the contour integral is equal to the *sum* of the contour integrals around the individual poles z_j. Use this to show that for this situation:

$$\oint_C h(z)\, dz = 2\pi i \sum_j \text{Res}\, h(z_j) \qquad \text{(counterclockwise direction).} \tag{17.16}$$

17.3 Solving integrals without knowing the primitive function

The residue theorem has some applications to integrals that do not contain a complex variable at all! As an example consider the integral

$$I = \int_{-\infty}^{\infty} \frac{1}{1+x^2} dx. \tag{17.17}$$

If you know that $1/(1+x^2)$ is the derivative of $\arctan x$ it is not difficult to solve this integral:

$$I = [\arctan x]_{-\infty}^{\infty} = \frac{\pi}{2} - \left(-\frac{\pi}{2}\right) = \pi. \tag{17.18}$$

Now suppose you did not know that $\arctan x$ is the primitive function of $1/(1+x^2)$. In that case you would be unable to see that the integral (17.17) is equal to π. Complex integration offers a way to obtain the value of this integral in a systematic fashion.

First note that the path of integration in the integral can be viewed as the real axis in the complex plane. Nothing prevents us from viewing the real function $1/(1+x^2)$ as a complex function $1/(1+z^2)$ because on the real axis z is equal to x. This means that

$$I = \int_{C_{real}} \frac{1}{1+z^2} dz, \tag{17.19}$$

where C_{real} denotes the real axis as integration path. At this point we cannot yet apply the residue theorem because the integration is not over a closed contour in the complex plane, see Figure 17.2. Let us close the integration path using the circular

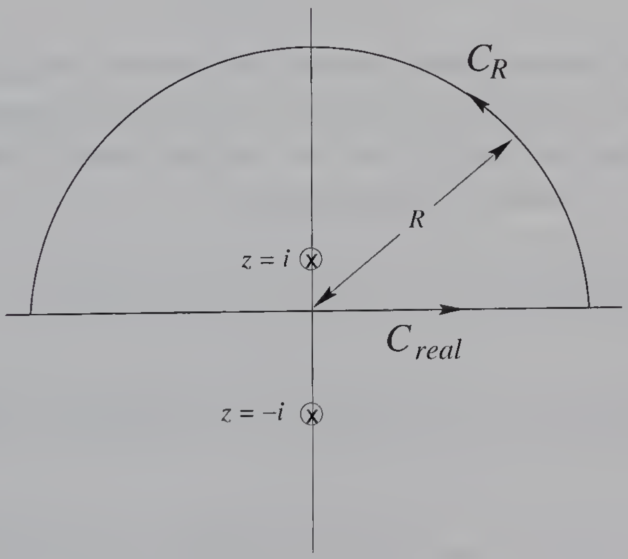

Fig. 17.2 Definition of the integration paths in the complex plane.

path C_R in the upper half-plane with radius R, see Figure 17.2. At the end of the calculation we will let R go to infinity so that the integration over the semicircle moves to infinity.

Problem a Show that

$$I = \int_{C_{real}} \frac{1}{1+z^2} dz = \oint_C \frac{1}{1+z^2} dz - \int_{C_R} \frac{1}{1+z^2} dz. \tag{17.20}$$

The circular integral is over the closed contour in Figure 17.2. What we have done is close the contour and subtract the integral over the semicircle that we added to obtain a closed integration path. This is the general approach in the application of the residue theorem; one adds segments to the integration path to obtain an integral over a closed contour in the complex plane, and corrects for the segments that one has added to the path. Obviously, this is only useful when the integral over the segment that one has added can be computed easily or when it vanishes. In this example the integral over the semicircle vanishes as $R \to \infty$. This can be seen from the following estimations:

$$\left| \int_{C_R} \frac{1}{1+z^2} dz \right| \leq \int_{C_R} \left| \frac{1}{1+z^2} \right| |dz| \leq \int_{C_R} \frac{1}{|z|^2 - 1} |dz| = \frac{\pi R}{R^2 - 1} \to 0$$
$$\text{as } R \to \infty. \tag{17.21}$$

The estimate (17.21) implies that the last integral in (17.20) vanishes in the limit $R \to \infty$. This means that

$$I = \oint_C \frac{1}{1+z^2} dz. \tag{17.22}$$

Now we are in the position to apply the residue theorem because we have reduced the problem to the evaluation of an integral along a closed contour in the complex plane. We know from Section 17.2 that this integral is determined by the poles of the function that is integrated within the contour.

Problem b Show that the function $1/(1+z^2)$ has poles for $z = +i$ and $z = -i$.

Only the pole at $z = +i$ is within contour C, see Figure 17.2. Since $1/(1+z^2) = 1/[(z-i)(z+i)]$ this pole is simple (why?).

Problem c Use (17.14) to show that for the pole at $z = i$ the residue is given by: Res $= 1/2i$.

Problem d Use the residue theorem (17.12) to deduce that

$$I = \int_{-\infty}^{\infty} \frac{1}{1+x^2} dx = \pi. \tag{17.23}$$

This value is identical to the value obtained at the beginning of this section by using that the primitive of $1/(1+x^2)$ is equal to $\arctan x$. Note that the analysis is systematic and that we did not need to "know" the primitive function.

In the treatment of this problem there is no reason why the contour should be closed in the upper half-plane. The estimate (17.21) holds equally well for a semicircle in the lower half-plane.

Problem e Repeat the analysis of this section with the contour closed in the lower half-plane. Use that now the pole at $z = -i$ contributes and take into account that the sense of integration is in the clockwise rather than the anticlockwise direction. Show that this leads to the same result (17.23) that was obtained by closing the contour in the upper half-plane.

In the evaluation of the integral (17.17) there was freedom whether to close the contour in the upper half-plane or in the lower half-plane. This is not always the case. To see this consider the integral

$$J = \int_{-\infty}^{\infty} \frac{\cos(x - \pi/4)}{1+x^2} dx. \tag{17.24}$$

Since $e^{ix} = \cos x + i \sin x$ this integral can be written as

$$J = \Re e \left(\int_{-\infty}^{\infty} \frac{e^{i(x-\pi/4)}}{1+x^2} dx \right), \tag{17.25}$$

where $\Re e(\cdots)$ again denotes the real part. We want to evaluate this integral by closing this integration path with a semicircle either in the upper half-plane or in the lower half-plane. Due to the term e^{ix} in the integral we now have no real choice in this issue. The decision whether to close the integral in the upper half-plane or in the lower half-plane is dictated by the requirement that the integral over the semicircle vanishes as $R \to \infty$. This can only happen when the integrand vanishes (faster than $1/R$) as $R \to \infty$. Let z be a point in the complex plane on the semicircle C_R that we use for closing the contour. On the semicircle z can be written as $z = Re^{i\varphi}$, where in the upper half-plane $0 \leq \varphi < \pi$ and in the lower half-plane $\pi \leq \varphi < 2\pi$.

Problem f Use this representation of z to show that

$$\left| e^{iz} \right| = e^{-R \sin\varphi}. \tag{17.26}$$

Problem g Show that in the limit $R \to \infty$ this term only goes to zero when z is in the upper half-plane.

This means that the integral over the semicircle only vanishes when we close the contour in the upper half-plane. Using steps similar to those in (17.21) one can show that the integral over a semicircle in the upper half-plane vanishes as $R \to \infty$.

Problem h Take exactly the same steps as in the derivation of (17.23) and show that

$$\int_{-\infty}^{\infty} \frac{\cos(x - \pi/4)}{1 + x^2} dx = \frac{\pi}{\sqrt{2}\, e}. \tag{17.27}$$

Note that this integral is equal to a combination of the three irrational numbers π, $\sqrt{2}$, and e!

Problem i Determine the integral $\int_{-\infty}^{\infty} [\sin(x - \pi/4)/(1 + x^2)] dx$ without doing any additional calculations. Hint: look carefully at (17.25) and spot $\sin(x - \pi/4)$.

It is important to note that in finding the integrals in this section we did not need to know the primitive function of the function that we integrated. This is the reason why complex integration is such a powerful technique; it allows us to solve many integrals without knowing their primitive functions. The structure of the singularities of the integrand in the complex plane and the behavior of the function on a bounding contour carry sufficient information to integrate the function. This rather miraculous property is due to the fact that according to (16.6) in regions where the integrand is analytic it satisfies Laplace's equation: $\nabla^2 f = \nabla^2 g = 0$ (where f and g are the real and imaginary parts of the integrand respectively). The singularities of the integrand act as sources or sinks for f and g. We have learned already that the specification of the sources and sinks (ρ) plus boundary conditions is sufficient to solve Poisson's equation: $\nabla^2 V = \rho$. In the same way, the specification of the singularities (the sources and sinks) of a complex function plus its values on a bounding contour (the boundary conditions) are sufficient to compute the integral of this complex function.

17.4 Response of a particle in syrup

Up to this point, contour integration has been applied to mathematical problems. However, this technique does have important applications in physical problems. In this section we consider a particle with mass m on which a force $f(t)$ is acting. The particle is suspended in syrup, which damps the velocity of the particle, and it is assumed that this damping force is proportional to the velocity $v(t)$ of the particle. The equation of motion of the particle is given by

$$m\frac{dv}{dt} + \beta v = f(t), \tag{17.28}$$

where β is a parameter that determines the strength of the damping of the motion by the fluid. The question we want to solve is: what is the velocity $v(t)$ for a given force $f(t)$?

We solve this problem using a Fourier transform technique. The Fourier transform of $v(t)$ is denoted by $V(\omega)$. The velocity in the frequency domain is related to the velocity in the time domain by the relation:

$$v(t) = \int_{-\infty}^{\infty} V(\omega)e^{-i\omega t}\,d\omega, \tag{17.29}$$

and its inverse

$$V(\omega) = \frac{1}{2\pi}\int_{-\infty}^{\infty} v(t)e^{+i\omega t}\,dt. \tag{17.30}$$

The force $f(t)$ is Fourier transformed using the same expressions; in the frequency domain it is denoted by $F(\omega)$.

Problem a Use the definitions of the Fourier transform to show that the equation of motion (17.28) is given in the frequency domain by

$$-i\omega m V(\omega) + \beta V(\omega) = F(\omega). \tag{17.31}$$

Comparing this with the original equation (17.28) we can see why the Fourier transform is so useful. The original expression (17.28) is a differential equation, while (17.31) is an algebraic equation. Since algebraic equations are much easier to solve than differential equations we have made considerable progress.

Problem b Solve the algebraic equation (17.31) for $V(\omega)$ and use the Fourier transform (17.29) to derive that

$$v(t) = \frac{i}{m}\int_{-\infty}^{\infty} \frac{F(\omega)e^{-i\omega t}}{\left(\omega + \dfrac{i\beta}{m}\right)}\,d\omega. \tag{17.32}$$

Now we have an explicit relation between the velocity $v(t)$ in the time domain and the force $F(\omega)$ in the frequency domain. This is not quite what we want since we want to find the relation between the velocity and the force $f(t)$ in the time domain.

Problem c Use the inverse Fourier transform (17.30) (but for the force) to show that

$$v(t) = \frac{i}{2\pi m} \int_{-\infty}^{\infty} f(t') \int_{-\infty}^{\infty} \frac{e^{-i\omega(t-t')}}{\left(\omega + \dfrac{i\beta}{m}\right)} \, d\omega \, dt' . \tag{17.33}$$

This equation looks messy, but we can simplify it by writing it as

$$v(t) = \frac{i}{2\pi m} \int_{-\infty}^{\infty} f(t') I(t-t') \, dt' , \tag{17.34}$$

with

$$I(t-t') = \int_{-\infty}^{\infty} \frac{e^{-i\omega(t-t')}}{\left(\omega + \dfrac{i\beta}{m}\right)} \, d\omega . \tag{17.35}$$

A comparison of (17.34) with expression (15.57) shows that $I(t-t')$ acts as a linear filter that converts the input $f(t)$ into the output $v(t)$.

The problem we now face is to evaluate this last integral. For this we use complex integration. The integration variable is now called ω rather than z but this does not change the principles. We close the contour by adding a semicircle either in the upper half-plane or in the lower half-plane to the integral (17.35) along the real axis. On a semicircle with radius R the complex number ω can be written as $\omega = Re^{i\varphi}$.

Problem d Show that $\left|e^{-i\omega(t-t')}\right| = e^{-R(t'-t)\sin\varphi}$.

Problem e The integral along the semicircle should vanish in the limit $R \to \infty$. Use the result of problem d to show that the contour should be closed in the upper half-plane for $t < t'$ and in the lower half-plane for $t > t'$, see Figure 17.3.

Problem f Show that the integrand in (17.35) has one pole at the negative imaginary axis at $\omega = -i\beta/m$ and that the residue at this pole is given by

$$Res = e^{-(\beta/m)(t-t')} . \tag{17.36}$$

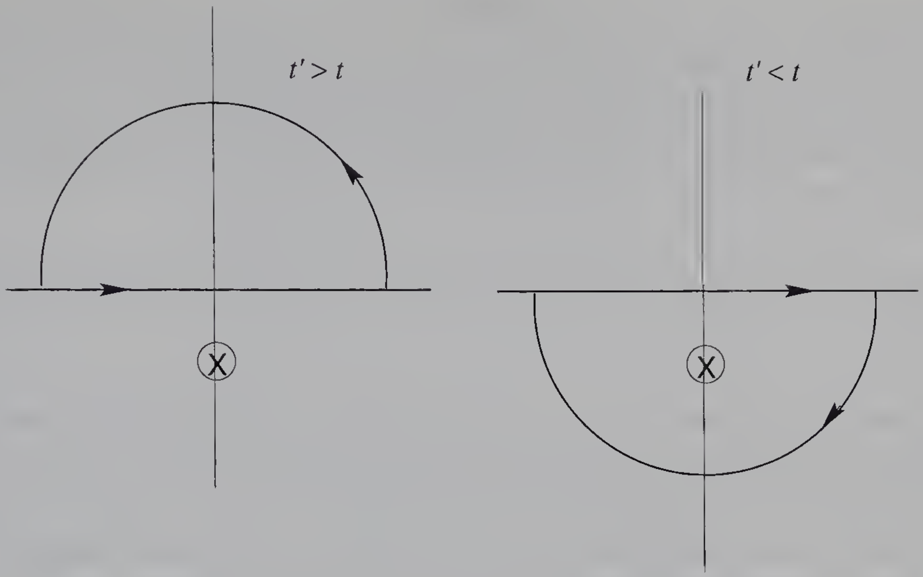

Fig. 17.3 The poles in the complex plane and the closure of contours for $t' > t$ (left) and $t' < t$ (right).

Problem g Use these results and the theorems derived in Section 17.2 to show that:

$$I(t-t') = \begin{cases} 0 & \text{for } \quad t < t' \\ -2\pi i e^{-(\beta/m)(t-t')} & \text{for } \quad t > t' \end{cases}. \tag{17.37}$$

Hint: treat the cases $t < t'$ and $t > t'$ separately.

Let us first consider (17.37). One can see from (17.33) that $I(t-t')$ is a function which describes the effect of a force acting at time t' on the velocity at time t. Expression (17.37) tells us that this effect is zero when $t < t'$. In other words, this expression tells us that the force $f(t')$ has no effect on the velocity $v(t)$ when $t < t'$. This is equivalent to saying that the force only affects the velocity at *later* times. In this sense, (17.37) can be seen as an expression of *causality*; the cause $f(t')$ only influences the effect $v(t)$ for later times.

Problem h Insert (17.37) and (17.35) in (17.33) to show that

$$v(t) = \frac{1}{m}\int_{-\infty}^{t} e^{-(\beta/m)(t-t')} f(t')\, dt'. \tag{17.38}$$

Pay particular attention to the limits of integration.

Problem i Convince yourself that this expression states that the force (the "cause") only has an influence on the velocity (the "effect") for *later* times.

It may appear that for this problem we have managed to give a proof of the causality principle. However, there is a problem hidden in the analysis. Suppose we switch off the damping parameter β, that is we remove the syrup from the problem. One can see that setting $\beta = 0$ in the final result (17.38) poses no problem. However, suppose that we had set $\beta = 0$ at the start of the problem.

Problem j Show that in that case the pole in Figure 17.3 is located on the *real* axis.

This implies that it is not clear how this pole affects the response. In particular, it is not clear whether this pole gives a nonzero contribution for $t < t'$ (as it would if we consider it to lie in the upper half-plane) or for $t > t'$ (as it would if we consider it to lie in the lower half-plane). This is a disconcerting result since it implies that causality only follows from the analysis when the problem contains some dissipation. This is not an artifact of the analysis using complex integration. What we are encountering here is a manifestation of the problem that in the absence of dissipation the laws of physics are symmetric for time-reversal, whereas the world around us seems to move in only one time direction This is the poorly resolved issue of the "arrow of time" [30, 83, 85, 100].

18

Green's functions: principles

Green's functions play an important role in mathematical physics. The Green's function plays a similar role to that of the impulse response for linear filters that was treated in Section 15.7. The general idea is that if one knows the response of a system to a delta-function input, the response of the system to any input can be reconstructed by superposing the response to the delta-function input in an appropriate manner. However, the use of Green's functions suffers from the same limitation as the use of the impulse response for linear filters: since the superposition principle underlies the use of Green's functions they are only useful for systems that are linear. Excellent treatments of Green's functions can be found in the book of Barton [9] which is completely devoted to Green's functions and also in Butkov [24]. Since the principles of Green's functions are so fundamental, the first section of this chapter is not presented as a set of problems.

18.1 Girl on a swing

In order to become familiar with Green's functions let us consider the example of a girl on a swing which is pushed by her mother, see Figure 18.1. When the amplitude of the swing is not too large, the motion of the swing is described by the equation of a harmonic oscillator that is driven by an external force $F(t)$ which is a general function of time:

$$\ddot{x} + \omega_0^2 x = F(t)/m. \tag{18.1}$$

The eigenfrequency of the oscillator is denoted by ω_0. It is not trivial to solve this equation for a general driving force. For simplicity, we solve (18.1) for the special case that the mother gives a single push to her daughter. The push is given at time $t = 0$, has duration Δ, and the magnitude of the force is denoted by F_0. This means

Fig. 18.1 The girl on a swing.

that:

$$F(t) = \begin{cases} 0 & \text{for} \quad t < 0 \\ F_0 & \text{for} \quad 0 \leq t < \Delta \, . \\ 0 & \text{for} \quad \Delta \leq t \end{cases} \tag{18.2}$$

We will look here for a *causal* solution. This is another way of saying that we are looking for a solution in which the cause (the driving force) precedes the effect (the motion of the oscillator). This means that we require that the oscillator does not move for times $t < 0$.

When $t < 0$ and $t \geq \Delta$ the function $x(t)$ satisfies the differential equation $\ddot{x} + \omega_0^2 x = 0$. This differential equation has general solutions of the form $x(t) = A \cos(\omega_0 t) + B \sin(\omega_0 t)$. For $t < 0$ the displacement vanishes, hence the constants A and B vanish. This determines the solution for $t < 0$ and $t \geq \Delta$. For $0 \leq t < \Delta$ the displacement satisfies the differential equation $\ddot{x} + \omega_0^2 x = F_0/m$. The solution can be found by writing $x(t) = F_0/m\omega_0^2 + y(t)$. The function $y(t)$ then satisfies the equation $\ddot{y} + \omega_0^2 y = 0$, which has the general solution $C \cos(\omega_0 t) + D \sin(\omega_0 t)$.

This means that the general solution $x(t)$ of the oscillator is given by

$$x(t)=\begin{cases}0 & \text{for} \quad t<0 \\ F_0/m\omega_0^2+C\cos(\omega_0 t)+D\sin(\omega_0 t) & \text{for} \quad 0\le t<\Delta, \\ A\cos(\omega_0 t)+B\sin(\omega_0 t) & \text{for} \quad \Delta\le t\end{cases} \tag{18.3}$$

where A, B, C, and D are integration constants which are not yet known.

These integration constants follow from the requirement that the displacement $x(t)$ of the oscillator is at all times continuous and that the velocity $\dot{x}(t)$ of the oscillator is at all times continuous. The last condition follows from the consideration that when the force is finite, the acceleration is finite and the velocity therefore is continuous. The requirement that both $x(t)$ and $\dot{x}(t)$ are continuous at $t=0$ and at $t=\Delta$ leads to the following equations:

$$\left.\begin{aligned} F_0/m\omega_0^2+C&=0, \\ \omega_0 D&=0, \\ F_0/m\omega_0^2+C\cos(\omega_0\Delta)+D\sin(\omega_0\Delta)&=A\cos(\omega_0\Delta)+B\sin(\omega_0\Delta), \\ -C\sin(\omega_0\Delta)+D\cos(\omega_0\Delta)&=-A\sin(\omega_0\Delta)+B\cos(\omega_0\Delta). \end{aligned}\right\} \tag{18.4}$$

These equations are four linear equations for the four unknown integration constants A, B, C, and D. The upper two equations can be solved directly for the constants C and D to give the values $C=-\left(F_0/m\omega_0^2\right)$ and $D=0$. These values for C and D can then be inserted into the lower two equations. Solving these equations for the constants A and B gives the values $A=-\left(F_0/m\omega_0^2\right)[1-\cos(\omega_0\Delta)]$ and $B=\left(F_0/m\omega_0^2\right)\sin(\omega_0\Delta)$. Inserting these values of the constants into (18.3) shows that the motion of the oscillator is given by:

$$x(t)=\begin{cases}0 & \text{for} \quad t<0 \\ (F_0/m\omega_0^2)[1-\cos(\omega_0 t)] & \text{for} \quad 0\le t<\Delta. \\ (F_0/m\omega_0^2)\{\cos[\omega_0(t-\Delta)]-\cos(\omega_0 t)\} & \text{for} \quad \Delta\le t\end{cases} \tag{18.5}$$

This is the solution for a push of duration Δ delivered at time $t=0$. Suppose now that the push is short compared to the period of the oscillator, then $\omega_0\Delta\ll 1$. In that case one can use a Taylor expansion in $\omega_0\Delta$ for the term $\cos[\omega_0(t-\Delta)]$ in (18.5). This can be achieved by using that $\cos[\omega_0(t-\Delta)]=\cos(\omega_0 t)\cos(\omega_0\Delta)-\sin(\omega_0 t)\sin(\omega_0\Delta)$ and by using the Taylor expansions $\sin x=x-x^3/6+O(x^5)$ and $\cos x=1-x^2/2+O(x^4)$ for $\sin(\omega_0\Delta)$ and $\cos(\omega_0\Delta)$. Retaining only the term of order $(\omega_0\Delta)$ and ignoring terms of higher order in $(\omega_0\Delta)$ shows that for an impulsive push $(\omega_0\Delta\ll 1)$ the solution is given by:

$$x(t)=\begin{cases}0 & \text{for} \quad t<0 \\ (F_0/m\omega_0^2)(\omega_0\Delta)\sin(\omega_0 t) & \text{for} \quad t>\Delta\end{cases}. \tag{18.6}$$

We will not bother anymore with the solution between $0 \leq t < \Delta$ because in the limit $\Delta \to 0$ this interval is of vanishing duration.

At this point we have all the ingredients needed to determine the response of the oscillator for a general driving force $F(t)$. Suppose we divide the time-axis in intervals of duration Δ. In the ith interval, the force is given by $F_i = F(t_i)$, where t_i is the time of the ith interval. We know from (18.6) the response to a force of duration Δ at time $t = 0$. The response to a force F_i at time t_i follows by replacing F_0 by F_i and by replacing t by $t - t_i$. Making these replacements it thus follows that the response to a force F_i delivered over a time interval Δ at time t_i is given by:

$$x(t) = \begin{cases} 0 & \text{for} \quad t < t_i \\ \dfrac{1}{m\omega_0} \sin\left[\omega_0 (t - t_i)\right] \; F(t_i)\Delta & \text{for} \quad t > t_i \end{cases} . \tag{18.7}$$

This is the response due to the force acting at time t_i only. To obtain the response to the full force $F(t)$ one should sum over the forces delivered at all times t_i. In the language of the girl on the swing one would say that (18.6) gives the motion of the swing for a single impulsive push, and that we want to use this result to find the displacement caused by a number of pushes. Since the differential equation (18.1) is linear we can use the superposition principle which states that the response to the superposition of two pushes is the sum of the response to the individual pushes. (In the language of Section 15.7 we would say that the swing is a linear system.) This means that when the swing receives a number of pushes at different times t_i the response can be written as the sum of the responses to every individual push. With (18.7) this gives:

$$x(t) = \sum_{t_i < t} \frac{1}{m\omega_0} \sin\left[\omega_0 (t - t_i)\right] \; F(t_i)\Delta. \tag{18.8}$$

Note that in (18.7) the response to a push at time t, before the push at time t_i, vanishes. For this reason one needs to sum in (18.8) only over the pushes at *earlier* times because the pushes at later times give a vanishing contribution. For this reason the summation is limited to times $t \geq t_i$.

Suppose now that the swing is not given a finite number of impulse pushes but instead that the driving force is a continuous function. This case can be handled by taking the limit $\Delta \to 0$. The summation in (18.8) then needs to be replaced by an integration. This can naturally be achieved because the duration Δ is then equal to the infinitesimal interval dt used in the integration. What we are really doing here is replacing the continuous function $F(t)$ by a function that is constant within every interval Δ at times t_i, see Figure 18.2, and then taking the limit where the width of the intervals goes to zero, that is $\Delta \to 0$. A similar treatment may be familiar to

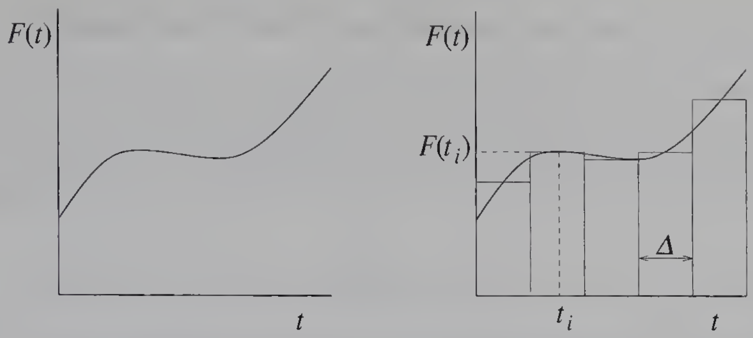

Fig. 18.2 A continuous function (left) and an approximation to this function that is constant within finite intervals (right).

you from the theory of integration. When the limit $\Delta \to 0$ is taken the summation over t_i can be replaced by an integration: $\sum_{t_i} (\cdots) \Delta \to \int (\cdots) d\tau$. The integration variable τ plays the role of the summation variable t_i and the time interval Δ is replaced $d\tau$. The response of the oscillator to a continuous force $F(t)$ is then given by

$$x(t) = \int_{-\infty}^{t} \frac{1}{m\omega_0} \sin\left[\omega_0 (t - \tau)\right] F(\tau)\, d\tau. \tag{18.9}$$

The integration is only carried out over times $\tau < t$ because the summation (18.8) extends only over the times $t_i < t$.

With a slight change in notation this result can be written as:

$$x(t) = \int_{-\infty}^{\infty} G(t, \tau) F(\tau)\, d\tau, \tag{18.10}$$

with

$$G(t, \tau) = \begin{cases} 0 & \text{for} \quad t < \tau \\ \dfrac{1}{m\omega_0} \sin\left[\omega_0 (t - \tau)\right] & \text{for} \quad t > \tau \end{cases}. \tag{18.11}$$

The function $G(t, \tau)$ in (18.11) is called the *Green's function* of the harmonic oscillator. Note that (18.10) is similar to the response of a linear filter described by (15.55). This is not surprising; in both examples the response of a linear system to an impulsive input was determined, and it will be no surprise that the results are identical. In fact, the Green's function is defined as the response of a linear system to a delta-function input. Although Green's functions are often presented in a rather abstract way, one should remember that:

> *The Green's function of a system is nothing but the impulse response of the system, that is it is the response of the system to a delta-function excitation.*

18.2 You have seen Green's functions before!

Although the concept of a Green's function may appear to be new to you, you have already seen several examples of Green's functions. One example is the electric field generated by a point charge q at the origin which was treated in Section 6.1:

$$\mathbf{E}(\mathbf{r}) = \frac{q\hat{\mathbf{r}}}{4\pi\varepsilon_0 r^2}. \tag{6.2}$$

Since this is the electric field generated by a delta-function charge at the origin, it is closely related to the Green's function for this problem. The field equation (6.23) of the electric field is invariant for translations in space. This is a complex way of saying that the electric field depends only on the *relative* positions of the point charge and the point of observation.

Problem a Show that this implies that the electric field at location $\mathbf{r}$ due to a point charge at location $\mathbf{r}'$ is given by:

$$\mathbf{E}(\mathbf{r}) = \frac{q}{4\pi\varepsilon_0}\frac{(\mathbf{r}-\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|^3}. \tag{18.12}$$

Now suppose that we do not have a single point charge, but instead a system of point charges q_i at locations $\mathbf{r}_i$. Since the field equation is linear, the electric field generated by a sum of point charges is the sum of the fields generated by each point charge:

$$\mathbf{E}(\mathbf{r}) = \sum_i \frac{q_i}{4\pi\varepsilon_0}\frac{(\mathbf{r}-\mathbf{r}_i)}{|\mathbf{r}-\mathbf{r}_i|^3}. \tag{18.13}$$

Problem b To which expression of the previous section does this equation correspond?

Just as in the previous section we now make the transition from a finite number of discrete inputs (either pushes of the swing or point charges) to an input function that is a continuous function (either the applied force to the oscillator as a function of time or a continuous electric charge). Let the electric charge per unit volume be denoted by $\rho(\mathbf{r})$. This means that the electric charge in a volume dV is given by $\rho(\mathbf{r})dV$.

Problem c Replace the sum in (18.13) by an integration over volume and use the appropriate charge for each volume element dV to show that the electric field for a continuous charge distribution is given by:

$$\mathbf{E}(\mathbf{r}) = \iiint \frac{\rho(\mathbf{r}')}{4\pi\varepsilon_0} \frac{(\mathbf{r} - \mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|^3} dV', \tag{18.14}$$

where the volume integration is over $\mathbf{r}'$.

Problem d Show that this implies that the electric field can be written as

$$\mathbf{E}(\mathbf{r}) = \iiint \mathbf{G}(\mathbf{r}, \mathbf{r}')\rho(\mathbf{r}')\, dV', \tag{18.15}$$

with the Green's function given by

$$\mathbf{G}(\mathbf{r}, \mathbf{r}') = \frac{1}{4\pi\varepsilon_0} \frac{(\mathbf{r} - \mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|^3}. \tag{18.16}$$

Note that this Green's function has the same form as the electric field for a point charge given in (18.12). Note that the Green's function is only a function of the relative distance $\mathbf{r} - \mathbf{r}'$.

Problem e Explain why the integral (18.15) can be seen as a three-dimensional convolution: $\mathbf{E}(\mathbf{r}) = \iiint \mathbf{G}(\mathbf{r} - \mathbf{r}')\rho(\mathbf{r}')dV'$.

The main purpose of this section is not to show you that you had seen an example of a Green's function before. Rather, it provides an example that the Green's function is not necessarily a function of time and that it is not necessarily a scalar function; the Green's function (18.16) depends on the position but not on time, and it describes a vector field rather than a scalar. The most important thing to remember is that the Green's function is the impulse response of a linear system.

Problem f You have seen another Green's function before if you worked through Section 17.4 where the response of a particle in syrup was treated. Find the Green's function in that section and spot the expressions equivalent to (18.10) and (18.11).

18.3 Green's function as impulse response

You may have found the derivation of the Green's function in Section 18.1 rather complex. The reason for this is that in (18.3) the motion of the swing was determined before the push ($t < 0$), during the push ($0 < t < \Delta$) and after the push ($t > \Delta$). The requirement that the displacement x and the velocity $\dot{x}$ were continuous then led to the system of equations (18.4) with four unknowns. However, in the end we took the limit $\Delta \to 0$ and did not use the solution for time $0 < t < \Delta$. This suggests that this method of solution is unnecessarily complicated. This is indeed the case. In this section an alternative derivation of the Green's function (18.11) is given which is based on the idea that the Green's function $G(t, \tau)$ describes the motion of the oscillator due to a delta-function force at time τ:

$$\ddot{G}(t, \tau) + \omega_0^2 G(t, \tau) = \frac{1}{m}\delta(t - \tau). \tag{18.17}$$

Problem a For $t \neq \tau$ the delta function vanishes and the right-hand side of this expression is equal to zero. We are looking for the *causal* Green's function, that is the solution where the cause (the force) precedes the effect (the motion of the oscillator). Show that these conditions imply that for $t \neq \tau$ the Green's function is given by:

$$G(t, \tau) = \begin{cases} 0 & \text{for} \quad t < \tau \\ A\cos[\omega_0(t - \tau)] + B\sin[\omega_0(t - \tau)] & \text{for} \quad t > \tau \end{cases}, \tag{18.18}$$

where A and B are unknown integration constants.

The integration constants follow from the conditions at $t = \tau$. Since we have two unknown parameters we need to impose two conditions. The first condition is that the displacement of the oscillator is continuous at $t = \tau$. If this were not the case the velocity of the oscillator would be infinite at that moment.

Problem b Show that the requirement of continuity of the Green's function at $t = \tau$ implies that $A = 0$.

The second condition requires more care. We derive the second condition first mathematically and then explore its physical meaning. The second condition follows by integrating (18.17) over t from $\tau - \varepsilon$ to $\tau + \varepsilon$ and by taking the limit $\varepsilon \downarrow 0$. Integrating (18.17) in this way gives:

$$\int_{\tau-\varepsilon}^{\tau+\varepsilon} \ddot{G}(t, \tau)\, dt + \omega_0^2 \int_{\tau-\varepsilon}^{\tau+\varepsilon} G(t, \tau)\, dt = \frac{1}{m}\int_{\tau-\varepsilon}^{\tau+\varepsilon} \delta(t - \tau)\, dt. \tag{18.19}$$

Problem c Show that the right-hand side of (18.19) is equal to $1/m$, regardless of the value of ε.

Problem d Show that the absolute value of the second term on the left-hand side of (18.19) is smaller than $2\varepsilon\omega_0^2 \max(G)$, where $\max(G)$ is the maximum of

G over the integration interval. Since the Green's function is finite this means that the middle term vanishes in the limit $\varepsilon \downarrow 0$.

Problem e Show that the first term on the left-hand side of (18.19) is equal to $\dot{G}(t = \tau + \varepsilon, \tau) - \dot{G}(t = \tau - \varepsilon, \tau)$. This quantity will be denoted by $[\dot{G}(t, \tau)]_{t=\tau-\varepsilon}^{t=\tau+\varepsilon}$.

Problem f Use this result to show that in the limit $\varepsilon \downarrow 0$ expression (18.19) gives:

$$\left[\dot{G}(t, \tau)\right]_{t=\tau-\varepsilon}^{t=\tau+\varepsilon} = \frac{1}{m}. \tag{18.20}$$

Problem g Show that this condition together with the continuity of G implies that the integration constants in (18.18) have the values $A = 0$ and $B = 1/m\omega_0$, that is that the Green's function is given by:

$$G(t, \tau) = \begin{cases} 0 & \text{for} \quad t < \tau \\ \dfrac{1}{m\omega_0} \sin\left[\omega_0 (t - \tau)\right] & \text{for} \quad t > \tau \end{cases}. \tag{18.21}$$

A comparison with (18.11) shows that the Green's function derived in this section is identical to the Green's function derived in Section 18.1. Note that the solution was obtained here without invoking the motion of the oscillator during the moment of excitation. This would have been difficult because the duration of the excitation (a delta function) is equal to zero, if it can be defined at all.

There is, however, something strange about the derivation in this section. In Section 18.1 the solution was found by requiring that the displacement x and its first derivative $\dot{x}$ were continuous at all times. As used in problem b the first condition is also met by the solution (18.21). However, the derivative $\dot{G}$ is *not* continuous at $t = \tau$.

Problem h Which of the equations that you derived above states that the first derivative is not continuous?

Problem i $G(t, \tau)$ denotes the displacement of the oscillator. Show that expression (18.20) states that the velocity of the oscillator changes discontinuously at $t = \tau$.

Problem j Give a physical reason why the velocity of the oscillator was continuous in the first part of Section 18.1 and why the velocity is discontinuous for the Green's function derived in this section. Hint: how large is the force needed to produce a finite jump in the velocity of a particle when the force is applied over a time-interval of length zero (the width of the delta-function excitation).

How can we reconcile this result with the solution obtained in Section 18.1?

Problem k Show that the change in the velocity in the solution $x(t)$ in (18.7) is proportional to $F(t_i)\Delta$, that is that

$$[\dot{x}]_{t_i-\varepsilon}^{t_i+\varepsilon} = \frac{1}{m} F(t_i)\Delta. \tag{18.22}$$

This means that the change in the velocity depends on the strength of the force times the duration of the force. The physical reason for this is that the change in the velocity depends on the integral of the force over time divided by the mass of the particle.

Problem l Derive this last statement directly from Newton's law ($F = ma$).

When the force is finite and when $\Delta \to 0$, the jump in the velocity is zero and the velocity is continuous. However, when the force is infinite (as is the case for a delta function), the jump in the velocity is nonzero and the velocity is discontinuous.

In many applications the Green's function is the solution of a differential equation with a delta function as the excitation. This implies that some derivative, or combination of derivatives, of the Green's function is equal to a delta function at the point (or time) of excitation. This usually has the effect that the Green's function or its derivative is not a continuous function. The delta function in the differential equation usually leads to a singularity in the Green's function or its derivative.

18.4 Green's function for a general problem

In this section, the theory of Green's functions is treated in a more abstract fashion. Every linear differential equation for a function u with a source term F can be written symbolically as:

$$Lu = F. \tag{18.23}$$

For example in (18.1) for the girl on the swing, u is the displacement $x(t)$ while L is a differential operator given by

$$L = m\frac{d^2}{dt^2} + m\omega_0^2, \tag{18.24}$$

where it is understood that a differential operator acts term by term on the function to the right of the operator.

Problem a Find the differential operator L and the source term F for the electric field treated in Section 18.2 from the field equation (6.23).

In the notation used in this section, the Green's function depends on the position vector $\mathbf{r}$, but the results derived here are equally valid for a Green's function that depends only on time or on position and time. In general, the differential equation (18.23) must be supplemented with boundary conditions to give a unique solution. In this section the position of the boundary is denoted by $\mathbf{r}_B$ and it is assumed that the function u has the value u_B at the boundary:

$$u(\mathbf{r}_B) = u_B. \tag{18.25}$$

Let us first find a single solution to the differential equation (18.23) without bothering about boundary conditions. We follow the same treatment as in Section 15.7 where in (15.54) the input of a linear function was written as a superposition of delta functions. In the same way, the source function can be written as:

$$F(\mathbf{r}) = \int \delta(\mathbf{r}-\mathbf{r}')F(\mathbf{r}')\,dV'. \tag{18.26}$$

This expression follows from the properties of the delta function. One can interpret this expression as an expansion of the function $F(\mathbf{r})$ in delta functions because the integral (18.26) describes a superposition of delta functions $\delta(\mathbf{r}-\mathbf{r}')$ centered at $\mathbf{r}=\mathbf{r}'$; each of these delta functions is given a weight $F(\mathbf{r}')$. We want to use a Green's function to construct a solution. The Green's function $G(\mathbf{r},\mathbf{r}')$ is the response at location $\mathbf{r}$ due to a delta-function source at location $\mathbf{r}'$, that is the Green's function satisfies:

$$LG(\mathbf{r},\mathbf{r}') = \delta(\mathbf{r}-\mathbf{r}'). \tag{18.27}$$

The response to the input $\delta(\mathbf{r}-\mathbf{r}')$ is given by $G(\mathbf{r},\mathbf{r}')$, and the source functions can be written as a superposition of these delta functions with weight $F(\mathbf{r}')$. This suggests that a solution of (18.23) is given by a superposition of Green's functions $G(\mathbf{r},\mathbf{r}')$, where each Green's function has the same weight factor as the delta function $\delta(\mathbf{r}-\mathbf{r}')$ in the expansion (18.26) of $F(\mathbf{r})$ in delta functions. This means that a solution of (18.23) is given by:

$$u_P(\mathbf{r}) = \int G(\mathbf{r},\mathbf{r}')F(\mathbf{r}')\,dV'. \tag{18.28}$$

Problem b If you worked through Section 15.7 discuss the relation between this expression and (15.55) for the output of a linear function.

It is crucial to understand at this point that we have used three steps to arrive at (18.28): (i) the source function is written as a superposition of delta functions, (ii) the response of the system to each delta-function input is defined, and (iii) the solution is written as the same superposition of Green's function as was used in the expansion of the source function in delta functions:

$$\begin{array}{llcl} \delta(\mathbf{r}-\mathbf{r}') \leftrightarrow & F(\mathbf{r}) & \overset{(i)}{=} & \int \delta(\mathbf{r}-\mathbf{r}')F(\mathbf{r}')\,dV' \\ \Downarrow (ii) & & & \Downarrow \\ G(\mathbf{r},\mathbf{r}') \leftrightarrow & u_P(\mathbf{r}) & \overset{(iii)}{=} & \int G(\mathbf{r},\mathbf{r}')F(\mathbf{r}')\,dV' \end{array} . \tag{18.29}$$

Problem c Although this reasoning may sound plausible, we have not proved that $u_P(\mathbf{r})$ in (18.28) actually is a solution of the differential equation (18.23). Give a proof that this is indeed the case by letting the operator L act on (18.28) and by using (18.27) for the Green's function. Hint: the operator L acts on $\mathbf{r}$ while the integration is over $\mathbf{r}'$, the operator can thus be taken inside the integral.

It should be noted that we have not solved our problem yet, because u_P does not necessarily satisfy the boundary conditions. In fact, (18.28) is just one of the many possible solutions to (18.23). It is a particular solution of the inhomogeneous equation (18.23), and this is the reason why the subscript P is used. Equation (18.23) is called an *inhomogeneous equation* because the right-hand side is nonzero. When the right-hand side is zero one speaks of the *homogeneous equation*. This implies that a solution u_H of the homogeneous equation satisfies

$$Lu_H = 0. \tag{18.30}$$

Problem d In general one can add a solution of the homogeneous equation (18.30) to a particular solution, and the result still satisfies the inhomogeneous equation (18.23). Give a proof of this statement by showing that the function $u = u_P + u_H$ is a solution of (18.23). In other words show that the general solution of (18.23) is given by:

$$u(\mathbf{r}) = u_H(\mathbf{r}) + \int G(\mathbf{r},\mathbf{r}')F(\mathbf{r}')\,dV'. \tag{18.31}$$

Problem e The problem is that we still need to enforce the boundary conditions (18.25). This can be achieved by requiring that the solution u_H satisfies specific boundary conditions at $\mathbf{r}_B$. Insert (18.31) into the boundary conditions (18.25) and show that the required solution u_H of the homogeneous equation must satisfy the following boundary conditions:

$$u_H(\mathbf{r}_B) = u_B(\mathbf{r}_B) - \int G(\mathbf{r}_B, \mathbf{r}')F(\mathbf{r}')\,dV'. \tag{18.32}$$

This is all we need to solve the problem. What we have shown is that:

> *The total solution* (18.31) *is given by the sum of the particular solution* (18.28) *plus a solution of the homogeneous equation* (18.30) *that satisfies the boundary condition* (18.32).

This construction may appear to be very complex to you. However, you should realize that the main complexity is the treatment of the boundary condition. In many problems, the boundary condition dictates that the function vanishes at the boundary ($u_B = 0$) and the Green's function also vanishes at the boundary. It follows from (18.31) that in that case the boundary condition for the homogeneous solution is $u_H(\mathbf{r}_B) = 0$. This boundary condition is satisfied by the solution $u_H(\mathbf{r}) = 0$, which implies that in that case one can dispense with the addition of u_H to the particular solution $u_P(\mathbf{r})$.

Problem f Suppose that the boundary conditions do not prescribe the value of the solution at the boundary but that instead of (18.25) the normal derivative of the solution is prescribed by the boundary conditions:

$$\frac{\partial u}{\partial n}(\mathbf{r}_B) = \hat{\mathbf{n}} \cdot \nabla u(\mathbf{r}_B) = w_B, \tag{18.33}$$

where $\hat{\mathbf{n}}$ is the unit vector perpendicular to the boundary. How should the theory in this section be modified to accommodate this boundary condition?

The theory in this section is rather abstract. In order to make the issues at stake more explicit the theory is applied in the next section to the calculation of the temperature in the Earth.

18.5 Radiogenic heating and the Earth's temperature

As an application of the use of the Green's function we consider in this section the calculation of the temperature in the Earth and specifically the effect of the decay of radioactive elements in the crust on the temperature in the Earth. Several radioactive elements such as U_{235} do not fit well in the lattice of mantle rocks. For this reason, these elements are expelled from the material in the Earth's mantle and

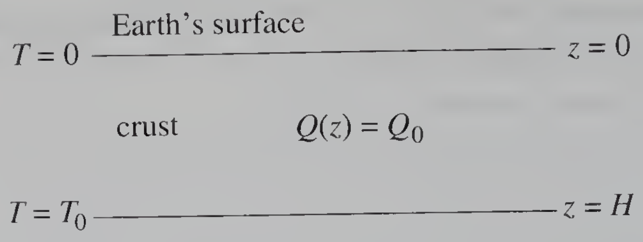

Fig. 18.3 Definition of the geometric variables and boundary conditions for the temperature in the crust.

they accumulate in the crust. Radioactive decay of these elements then leads to the production of heat at the place where these elements have accumulated.

As a simplified example of this problem we assume that the temperature T and the radiogenic heating Q depend on depth only, and that we can ignore the sphericity of the Earth. In addition, we assume that the radiogenic heating does not depend on time and that we consider only the equilibrium temperature.

Problem a Show that these assumptions imply that the temperature is only a function of the z-coordinate: $T = T(z)$.

The temperature field satisfies the heat equation derived in Section 11.4:

$$\frac{\partial T}{\partial t} = \kappa \nabla^2 T + Q. \tag{11.31}$$

Problem b Use this expression to show that for the problem in this section the temperature field satisfies

$$\frac{d^2 T}{dz^2} = -\frac{Q(z)}{\kappa}. \tag{18.34}$$

This equation can be solved when the boundary conditions are specified. The thickness of the crust is denoted by H, see Figure 18.3. The temperature is assumed to vanish at the Earth's surface. In addition, it is assumed that at the base of the crust the temperature has a fixed value T_0.[‡] This implies that the boundary conditions are:

$$T(z = 0) = 0, \qquad T(Z = H) = T_0. \tag{18.35}$$

In this section we solve the differential equation (18.34) with the boundary conditions (18.35) using the Green's function technique described in the previous section. Analogously to (18.28) we first determine a particular solution T_P of the differential equation (18.34) and worry about the boundary conditions later. The Green's function $G(z, z')$ to be used is the temperature at depth z due to delta-function heating at depth z':

[‡] Geophysically this is an oversimplified boundary condition because in reality the temperature in the Earth is determined by the radiogenic heating everywhere in the Earth and by the heat that was formed during the Earth's formation.

$$\frac{d^2G(z, z')}{dz^2} = \delta(z - z'). \tag{18.36}$$

Problem c Use the theory of the previous section to show that the following function satisfies the heat equation (18.34):

$$T_P(z) = -\frac{1}{\kappa}\int_0^H G(z, z')Q(z')\,dz'. \tag{18.37}$$

Before further progress can be made it is necessary to find the Green's function, that is to solve the differential equation (18.36). In order to do this the boundary conditions for the Green's function need to be specified. In this example we will use a Green's function that vanishes at the endpoints of the depth interval:

$$G(z = 0, z') = G(Z = H, z') = 0. \tag{18.38}$$

Problem d Use (18.36) to show that for $z \neq z'$ the Green's function satisfies the differential equation $d^2G(z, z')/dz^2 = 0$ and use this to show that the Green's function which satisfies the boundary conditions (18.38) must be of the form

$$G(z, z') = \begin{cases} \beta z & \text{for} \quad z < z' \\ \gamma(z - H) & \text{for} \quad z > z' \end{cases}, \tag{18.39}$$

with β and γ constants that need to be determined. These constants are independent of z, but they may depend on z'.

Problem e Since there are two unknown constants, two conditions are needed. The first condition is that the Green's function is continuous for $z = z'$. Use the theory of Section 18.3 and the differential equation (18.36) to show that the second requirement is:

$$\lim_{\varepsilon \downarrow 0}\left[\frac{dG(z, z')}{dz}\right]_{z=z'-\varepsilon}^{z=z'+\varepsilon} = 1, \tag{18.40}$$

that is that the first derivative makes a unit jump at the point of excitation.

Problem f Apply these two conditions to the solution (18.39) to determine the constants β and γ and show that the Green's function is given by:

$$G(z, z') = \begin{cases} -\dfrac{1}{H}(H - z')z & \text{for} \quad z < z' \\ \\ -\dfrac{1}{H}z'(H - z) & \text{for} \quad z > z' \end{cases} . \tag{18.41}$$

In this notation the two regions $z < z'$ and $z > z'$ are separated. Note, however, that the solution in the two regions has a highly symmetric form. In the literature you will find that a solution such as (18.41) is often rewritten by defining $z_>$ to be the maximum of z and z' and $z_<$ to be the minimum of z and z':

$$\left. \begin{aligned} z_> &\equiv \max(z, z'), \\ z_< &\equiv \min(z, z'). \end{aligned} \right\} \tag{18.42}$$

Problem g Show that in this notation the Green's function (18.41) can be written as:

$$G(z, z') = -\frac{1}{H}(H - z_>)z_<. \tag{18.43}$$

For simplicity we assume that the radiogenic heating is constant in the crust:

$$Q(z) = Q_0 \qquad \text{for} \quad 0 < z < H. \tag{18.44}$$

Problem h Show that the particular solution (18.37) for this heating function is given by

$$T_P(z) = \frac{Q_0 H^2}{2\kappa} \frac{z}{H} \left(1 - \frac{z}{H}\right). \tag{18.45}$$

Problem i Show that this particular solution satisfies the boundary conditions

$$T_P(z = 0) = T_P(z = H) = 0. \tag{18.46}$$

Problem j This means that this solution does not satisfy the boundary conditions (18.35) of our problem. Use the theory in Section 18.4 to show that to obtain this solution we must add a solution $T_H(z)$ of the homogeneous equation $d^2T_H/dz^2 = 0$ which satisfies the boundary conditions $T_H(z = 0) = 0$ and $T_H(z = H) = T_0$.

Problem k Show that the solution of the homogeneous equation is given by $T_H(z) = T_0 z/H$ and that the total solution is given by

$$T(z) = T_0 \frac{z}{H} + \frac{Q_0 H^2}{2\kappa} \frac{z}{H} \left(1 - \frac{z}{H}\right). \tag{18.47}$$

Problem l Verify explicitly that this solution satisfies the differential equation (18.34) with the boundary conditions (18.35).

As shown in expression (11.29) the conductive heat flow is given by $\mathbf{J} = -\kappa \nabla T$. Since the problem is one-dimensional the heat flow is given by

$$J = -\kappa \frac{dT}{dz}. \tag{18.48}$$

Problem m Compute the heat flow at the top ($z = 0$) and at the bottom ($z = H$) of the crust. Assuming that T_0 and Q_0 are both positive, does the heat flow at these locations increase or decrease because of the radiogenic heating Q_0? Give a physical interpretation of this result.

The derivation in this section used a Green's function which satisfies the boundary conditions (18.38) rather than the boundary conditions (18.35) of the temperature field. However, there is no particular reason why one should use these boundary conditions for the Green's function. That is to say, one might think one could avoid the step of adding a solution $T_H(z)$ of the homogeneous equation by using a Green's function $\tilde{G}$ that satisfies the differential equation (18.39) and an inhomogeneous boundary condition at the base of the crust:

$$\tilde{G}(z = 0, z') = 0, \qquad \tilde{G}(z = H, z') = H. \tag{18.49}$$

(The boundary value at the base of the crust is set equal to H because the Green's function has the dimension of length (see (18.41)) and the crustal thickness H is the only length-scale in the problem.)

Problem n Go through the same steps as you did earlier in this section by constructing the Green's function $\tilde{G}(z, z')$, computing the corresponding particular solution $\tilde{T}_P(z)$, verifying whether the boundary conditions (18.35) are satisfied by this particular solution and if necessary adding a solution of the homogeneous equation in order to satisfy the boundary conditions. Show that this again leads to (18.47).

The lesson to be learned from this section is that usually one needs to add a solution of the homogeneous equation to a particular solution in order to satisfy the boundary conditions. However, suppose that the boundary conditions of the temperature field were also homogeneous: $T = (z = 0) = T(z = H) = 0$. In that case the particular solution (18.45) that was constructed using a Green's function that satisfies the homogeneous boundary conditions (18.38) satisfies the boundary conditions of the full problem as well. This implies that it only pays to use a Green's function that satisfies the boundary conditions of the full problem when these boundary conditions are *homogeneous*, that is when the function itself vanishes ($T = 0$) or when the normal gradient of the function vanishes ($\partial T/\partial n = 0$) or when a linear combination of these quantities vanishes ($aT + b\partial T/\partial n = 0$). In all other cases one cannot avoid adding a solution of the homogeneous equation in order to satisfy the boundary conditions and the most efficient procedure is usually to use the Green's function that can most easily be computed.

18.6 Nonlinear systems and Green's functions

Up to this point, Green's functions have been applied to linear systems. The definition of a linear system was introduced in Section 15.7. Suppose that a force F_1 leads to a response x_1 and that a force F_2 leads to a response x_2. A system is linear when the response to the linear combination $c_1F_1 + c_2F_2$ (with c_1 and c_2 constants) is the superposition response $c_1x_1 + c_2x_2$. This definition implies that the response to the input times a constant is given by the response multiplied by the same constant. In other words show that for a linear system an input that is twice as large leads to a response that is twice as large.

Problem a Show that the definition of linearity given above implies that the response to the sum of two force functions is the sum of the responses to the individual force functions.

This last property reflects that a linear system satisfies the *superposition principle* which states that for a linear system one can superpose the response to a sum of force functions.

Not every system is linear, and we exploit here the extent to which Green's functions are useful for nonlinear systems. As an example we consider the Verhulst equation:

$$\dot{x} = x - x^2 + F(t). \tag{18.50}$$

This equation has been used in mathematical biology to describe the growth of a population. Suppose that only the term x were present on the right-hand side. In that case the solution would be given by $x(t) = Ce^t$. This means that the first term on the right-hand side accounts for the exponential population growth which is due to the fact that the number of offspring is proportional to the size of the population. However, a population cannot grow indefinitely; when a population is too large limited resources restrict the growth, and this is accounted for by the $-x^2$ term on the right-hand side. The term $F(t)$ accounts for external influences on the population. For example, a mass-extinction could be described by a strongly negative forcing function $F(t)$. We will consider first the solution for the case that $F(t) = 0$. Since the population size is positive we consider positive solutions $x(t)$ only.

Problem b Show that for the case $F(t) = 0$ the change of variable $y = 1/x$ leads to the linear equation $\dot{y} = 1 - y$. Solve this equation and show that the general solution of (18.50) (with $F(t) = 0$) is given by:

$$x(t) = \frac{1}{Ae^{-t} + 1}, \tag{18.51}$$

with A an integration constant.

Problem c Use this solution to show that any solution of the unforced equation goes to 1 for infinite times:

$$\lim_{t\to\infty} x(t) = 1. \tag{18.52}$$

In other words, the population of the unforced Verhulst equation always converges to the same population size. Note that when the force vanishes after a finite time, the solution after that time must satisfy (18.51) which implies that the long-time limit is then also given by (18.52).

Now, consider the response to a delta-function excitation at time t_0 with strength F_0. The associated response $g(t, t_0)$ thus satisfies

$$\dot{g} - g + g^2 = F_0\,\delta(t - t_0). \tag{18.53}$$

Since this function is the impulse response of the system, the notation g is used in order to bring out the resemblance with the Green's functions used earlier. We consider only causal solutions, that is we require that $g(t, t_0)$ vanishes for $t < t_0$: $g(t, t_0) = 0$ for $t < t_0$. For $t > t_0$ the solution satisfies the Verhulst equation without force, hence the general form is given by (18.51). The only task remaining is to find the integration constant A. This constant follows by a treatment similar to the analysis of Section 18.3.

Problem d Integrate (18.53) over t from $t_0 - \varepsilon$ to $t_0 + \varepsilon$, take the limit $\varepsilon \downarrow 0$ and show that this leads to the following requirement for the discontinuity in g:

$$\lim_{\varepsilon \downarrow 0} \left[g(t, t_0) \right]_{t_0-\varepsilon}^{t_0+\varepsilon} = F_0. \tag{18.54}$$

Problem e Use this condition to show that the constant A in the solution (18.51) is given by $A = (1/F_0 - 1)e^{t_0}$ and that the solution is given by:

$$g(t, t_0) = \begin{cases} 0 & \text{for} \quad t < t_0 \\ \dfrac{F_0}{(1 - F_0)e^{-(t-t_0)} + F_0} & \text{for} \quad t > t_0 \end{cases}. \tag{18.55}$$

At this point you should be suspicious of interpreting $g(t, t_0)$ as a Green's function. An important property of linear systems is that the response is proportional to the force. However, the solution $g(t, t_0)$ in (18.55) is not proportional to the strength F_0 of the force.

Let us now check whether we can use the superposition principle. Suppose the force function is the superposition of a delta-function force F_1 at $t = t_1$ and a delta-function force F_2 at $t = t_2$:

$$F(t) = F_1 \delta(t - t_1) + F_2 \delta(t - t_2). \tag{18.56}$$

By analogy with (18.10) you might think that a Green's function-type solution is given by:

$$x^{Green}(t) = \frac{F_1}{(1 - F_1)e^{-(t-t_1)} + F_1} + \frac{F_2}{(1 - F_2)e^{-(t-t_2)} + F_2}, \tag{18.57}$$

for times larger than both t_1 and t_2. You can verify by direct substitution that this function is not a solution of the differential equation (18.50). However, this process is rather tedious and there is a simpler way to see that the function $x^{Green}(t)$ violates the differential equation (18.50).

Problem f To see this, show that the solution $x^{Green}(t)$ has the following long-time behavior:

$$\lim_{t \to \infty} x^{Green}(t) = 2. \tag{18.58}$$

This limit is at odds with the limit (18.52) that every solution of the differential equation (18.50) should satisfy when the force vanishes after a certain finite time. This proves that $x^{Green}(t)$ is not a solution of the Verhulst equation.

This implies that the Green's function technique introduced in the previous sections cannot be used for a nonlinear equation such as the forced Verhulst equation. The reason for this is that Green's functions are based on the superposition principle; by knowing the response to a delta-function force and by writing a general force as a superposition of delta functions one can construct a solution by making the corresponding superposition of Green's functions, see (18.29). However, solutions of a nonlinear equation such as the Verhulst equation do not satisfy the principle of superposition. This implies that Green's function cannot be used effectively to construct the behavior of nonlinear systems. It is for this reason that Green's functions are in practice only used for constructing the response of *linear* systems.

19

Green's functions: examples

In the previous chapter the basic theory of Green's function was introduced. In this chapter a number of examples of Green's functions are shown that are often used in mathematical physics.

19.1 Heat equation in N dimensions

In this section we consider once again the heat equation as introduced in Section 11.4:

$$\frac{\partial T}{\partial t} = \kappa \nabla^2 T + Q. \tag{11.31}$$

First we construct a Green's function for this equation in N space dimensions. The reason for this is that the analysis for N dimensions is just as easy (or difficult) as the analysis for only one spatial dimension.

The heat equation is invariant for translations in both space and time. For this reason the Green's function $G(\mathbf{r},t;\mathbf{r}_0,t_0)$ that gives the temperature at location $\mathbf{r}$ and time t due to a delta-function heat source at location $\mathbf{r}_0$ and time t_0 depends only on the relative distance $\mathbf{r}-\mathbf{r}_0$ and the relative time $t-t_0$. This implies that $G(\mathbf{r},t;\mathbf{r}_0,t_0) = G(\mathbf{r}-\mathbf{r}_0,t-t_0)$. Since the Green's function depends only on $\mathbf{r}-\mathbf{r}_0$ and $t-t_0$ it suffices to construct the simplest solution by considering the special case of a source at $\mathbf{r}_0 = 0$ at time $t_0 = 0$. This means that we construct the Green's function $G(\mathbf{r},t)$ that satisfies:

$$\frac{\partial G(\mathbf{r},t)}{\partial t} - \kappa \nabla^2 G(\mathbf{r},t) = \delta(\mathbf{r})\delta(t). \tag{19.1}$$

This Green's function can most easily be constructed by carrying out a spatial Fourier transform. Using the Fourier transform (15.27) for each of the N spatial dimensions one finds that the Green's function has the following Fourier expansion:

$$G(\mathbf{r}, t) = \frac{1}{(2\pi)^N} \int g(\mathbf{k}, t) e^{i\mathbf{k}\cdot\mathbf{r}} d^N k. \tag{19.2}$$

Note that the Fourier transform is only carried out over the spatial dimensions and not over time. This implies that $g(\mathbf{k}, t)$ is a function of time as well. The differential equation that g satisfies can be obtained by inserting the Fourier representation (19.2) in the differential equation (19.1). In doing this we also need the Fourier representation of $\nabla^2 G(\mathbf{r},t)$.

Problem a Show by applying the Laplacian to the Fourier integral (19.1) that:

$$\nabla^2 G(\mathbf{r}, t) = \frac{-1}{(2\pi)^N} \int k^2 g(\mathbf{k}, t) e^{i\mathbf{k}\cdot\mathbf{r}} d^N k. \tag{19.3}$$

Problem b As a last ingredient we need the Fourier representation of the delta function on the right-hand side of (19.1). This multi-dimensional delta function is a shorthand notation for $\delta(\mathbf{r}) = \delta(x_1)\delta(x_2)\cdots\delta(x_N)$. Use the Fourier representation (15.31) of the delta function to show that:

$$\delta(\mathbf{r}) = \frac{1}{(2\pi)^N} \int e^{i\mathbf{k}\cdot\mathbf{r}} d^N k. \tag{19.4}$$

Problem c Insert these results into the differential equation (19.1) of the Green's function to show that $g(\mathbf{k}, t)$ satisfies the differential equation

$$\frac{\partial g(\mathbf{k}, t)}{\partial t} + \kappa k^2 g(\mathbf{k}, t) = \delta(t). \tag{19.5}$$

We have made considerable progress. The original equation (19.1) was a partial differential equation, whereas (19.5) is an ordinary differential equation for g because only a time derivative is taken. In fact, you saw this equation before when you read Section 17.4 which dealt with the response of a particle in syrup. Equation (19.5) is equivalent to the equation of motion (17.28) for a particle in syrup when the forcing force is a delta function.

Problem d Use the theory of Section 18.3 to show that the causal solution of (19.5) is given by:

$$g(\mathbf{k}, t) = e^{-\kappa k^2 t}. \tag{19.6}$$

This solution can be inserted into the Fourier representation (19.2) of the Green's function, and this gives:

$$G(\mathbf{r}, t) = \frac{1}{(2\pi)^N} \int e^{-\kappa k^2 t + i\mathbf{k}\cdot\mathbf{r}} d^N k. \tag{19.7}$$

The Green's function can be found by solving this Fourier integral. Before we do this, let us pause to consider the solution (19.6) for the Green's function in the wavenumber–time domain. The function $g(\mathbf{k}, t)$ gives the coefficient of the plane wave component $e^{i\mathbf{k}\cdot\mathbf{r}}$ as a function of time. According to (19.6) each Fourier component decays exponentially with time with a characteristic decay time $1/\kappa k^2$.

Problem e Show that this implies that in the Fourier expansion (19.2) plane waves with a smaller wavelength decay faster with time than plane waves with a larger wavelength. Explain this result physically.

In order to find the Green's function, we need to solve the Fourier integral (19.7). The integrations over the different components k_i of the wavenumber integration all have the same form.

Problem f Show this by giving a proof that the Green's function can be written as:

$$G(\mathbf{r}, t) = \frac{1}{(2\pi)^N} \left(\int e^{-\kappa k_1^2 t + i k_1 x_1} dk_1 \right) \left(\int e^{-\kappa k_2^2 t + i k_2 x_2} dk_2 \right) \times \cdots \left(\int e^{-\kappa k_N^2 t + i k_N x_N} dk_N \right). \tag{19.8}$$

You will notice that each of the integrals is of the same form, hence the Green's function can be written as

$$G(x_1, x_2, \ldots, x_N, t) = I(x_1, t) I(x_2, t) \cdots I(x_N, t)$$

with $I(x, t)$ given by

$$I(x, t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-\kappa k^2 t + ikx} dk. \tag{19.9}$$

This means that our problem is solved when the one-dimensional Fourier integral (19.9) is solved. In order to solve this integral it is important to realize that the exponent in the integral is a quadratic function of the integration variable k. If the integral were of the form $\int_{-\infty}^{\infty} e^{-\alpha k^2} dk$ the problem would not be difficult because

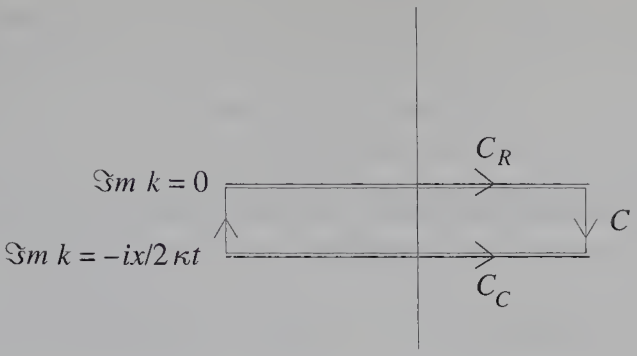

Fig. 19.1 The contours C_R, C_C, and C in the complex k-plane.

it is known that this integral has the value $\sqrt{\pi/\alpha}$. The problem can be solved by rewriting the integral (19.9) in the form of the integral $\int_{-\infty}^{\infty} e^{-\alpha k^2} dk$.

Problem g Complete the square of the exponent in (19.9), that is show that

$$-\kappa k^2 t + ikx = -\kappa t \left(k - \frac{ix}{2\kappa t} \right)^2 - \frac{x^2}{4\kappa t}, \tag{19.10}$$

and use this result to show that $I(x,t)$ can be written as:

$$I(x,t) = \frac{1}{2\pi} e^{-x^2/4\kappa t} \int_{-\infty - ix/2\kappa t}^{\infty - ix/2\kappa t} e^{-\kappa k'^2 t} dk'. \tag{19.11}$$

With these steps we have achieved our goal of having an integrand of the form $e^{-\alpha k^2}$, but have paid a price. In the integral (19.9) the integration was along the real axis C_R, see Figure 19.1. In the transformed integral the integration now takes place along the integration path C_C in the complex plane that lies below the real axis, see Figure 19.1. However, one can show that when the integration path C_C is replaced by an integration along the real axis the integral has the same value:

$$I(x,t) = \frac{1}{2\pi} e^{-x^2/4\kappa t} \int_{-\infty}^{\infty} e^{-\kappa k^2 t} dk. \tag{19.12}$$

Problem h When you studied Section 17.2 you saw all the material necessary to give a proof that (19.12) is indeed identical to (19.11). Show that this is indeed the case by using that the closed integral along the closed contour C in Figure 19.1 vanishes.

Problem i Carry out the integration in (19.12) and derive that

$$I(x,t) = \frac{e^{-x^2/4\kappa t}}{\sqrt{4\pi\kappa t}}, \tag{19.13}$$

and show that this implies that the Green's function is given by

$$G(\mathbf{r},t)=\frac{1}{(4\pi\kappa t)^{N/2}}e^{-r^2/4\kappa t}. \tag{19.14}$$

Problem j This result implies that the Green's function in any dimension has the form of the Gaussian. Show that this Gaussian changes shape with time. Is the Gaussian broadest at early times or at late times? What is the shape of the Green's function in the limit $t \downarrow 0$, that is at the time just after the heat force has been applied.

Problem k Sketch the time behavior of the Green's function for a fixed distance r. Does the Green's function decay more rapidly as a function of time in three dimensions than in one dimension? Give a physical interpretation of this result.

In one dimension, the Green's function (19.14) is given by

$$G(x,t)=\frac{1}{\sqrt{4\pi\kappa t}}e^{-x^2/4\kappa t}. \tag{19.15}$$

Problem l Show that this is the same function as the solution (11.41) for the heat equation in one dimension when the limits $L \to 0$ and $T_0 L \to 1$ are taken.

These limits to the Gaussian function define the delta function, see Section 14.1. This explains why in this limit the Green's function (19.15) and solution (11.41) are identical.

It is a remarkable property of the derivation in this section that the Green's function can be derived with a single derivation for every number of dimensions. It should be noted that this is not the generic case. In many problems, the behavior of the system depends critically on the number of spatial dimensions. We will see in Section 19.4 that wave propagation in two dimensions is fundamentally different from wave propagation in one or three dimensions. Another example is chaotic behavior of dynamical systems where the occurrence of chaos is intricately linked to the number of dimensions, see the discussion given by Tabor [104].

19.2 Schrödinger equation with an impulsive source

In this section we study the Green's function for the Schrödinger equation which was introduced in Section 8.4:

$$i\hbar\frac{\partial\psi(\mathbf{r},t)}{\partial t}=-\frac{\hbar^2}{2m}\nabla^2\psi(\mathbf{r},t)+V(\mathbf{r})\psi(\mathbf{r},t). \tag{8.14}$$

Solving this equation for a general potential $V(\mathbf{r})$ is a formidable problem, and solutions are known for only very few examples such as the free particle, the harmonic oscillator, and the Coulomb potential. We restrict ourselves here to the simplest case of a free particle: this is the case where the potential vanishes ($V(\mathbf{r}) = 0$). The corresponding Green's function satisfies the following partial differential equation:

$$\frac{\hbar}{i}\frac{\partial G(\mathbf{r},t)}{\partial t} - \frac{\hbar^2}{2m}\nabla^2 G(\mathbf{r},t) = \delta(\mathbf{r})\delta(t). \tag{19.16}$$

Before we compute the Green's function for this problem, let us pause to consider the meaning of this Green's function. First, the Green's function is a solution of Schrödinger's equation for $\mathbf{r} \neq 0$ and $t \neq 0$. This means that $|G|^2$ gives the probability density of a particle (see also Section 8.4). However, the right-hand side of (19.16) contains a delta-function forcing at time $t = 0$ at location $\mathbf{r} = 0$. This is a source term of G and hence this is a source of the probability of the presence of the particle. One can say that this source term creates the probability for having a particle at the origin at $t = 0$. Of course, this particle will not necessarily remain at the origin, it will move according to the laws of quantum mechanics. This motion is described by (19.16). This means that this equation describes the time evolution of matter waves when matter is injected at $t = 0$ at location $\mathbf{r} = 0$.

Problem a The Green's function $G(\mathbf{r}, t; \mathbf{r}', t')$ gives the wavefunction at location $\mathbf{r}$ and time t for a source of particles at location $\mathbf{r}'$ at time t'. Express the Green's function $G(\mathbf{r}, t; \mathbf{r}', t')$ in terms of the solution $G(\mathbf{r}, t)$ of (19.16), and show how you obtained this result. Is this result also valid for the Green's function for the quantum-mechanical harmonic oscillator (where the potential $V(\mathbf{r})$ depends on position)?

In the previous section the Green's function gave the evolution of the temperature field due to a delta-function injection of heat at the origin at time $t = 0$. Similarly, the Green's function of this section describes the time evolution of the probability of a delta-function injection of matter waves at the origin at time $t = 0$. These two Green's functions are not only conceptually very similar, the differential equations (19.1) for the temperature field and (19.16) for the Schrödinger equation are first order differential equations in time and second order differential equations in the space coordinate that have a delta-function excitation on the right-hand side. In this section we exploit this similarity and derive the Green's function for Schrödinger's equation from the Green's function for the heat equation derived in the previous section rather than constructing the solution from first principles. This approach is admittedly not very rigorous, but it shows that analogies are useful for making shortcuts.

The principle difference between (19.1) and (19.16) is that the time derivative for Schrödinger's equation is multiplied by $i = \sqrt{-1}$ whereas the heat equation is purely real. We will relate the two equations by introducing the new time variable τ for the Schrödinger equation that is proportional to the original time: $\tau = \gamma t$.

Problem b How should the proportionality constant γ be chosen so that (19.16) transforms to:

$$\frac{\partial G(\mathbf{r}, \tau)}{\partial \tau} - \frac{\hbar^2}{2m}\nabla^2 G(\mathbf{r}, \tau) = C\delta(\mathbf{r})\delta(\tau). \tag{19.17}$$

The constant C on the right-hand side cannot easily be determined from the change of variables $\tau = \gamma t$ because γ is not necessarily real and it is not clear how a delta function with a complex argument should be interpreted. For this reason we will not bother to specify C.

The key point to note is that this equation is of exactly the same form as the heat equation (19.1), where $\hbar^2/2m$ plays the role of the heat conductivity κ. The only difference is the constant C on the right-hand side of (19.17). However, since the equation is linear, this term only leads to an overall multiplication by C.

Problem c Show that the Green's function defined in (19.17) for the Schrödinger equation can be obtained from the Green's function (19.14) for the heat equation by making the following substitutions:

$$\left.\begin{array}{l} t \longrightarrow it/\hbar, \\ \kappa \longrightarrow \hbar^2/2m, \\ G \longrightarrow C\,G. \end{array}\right\} \tag{19.18}$$

It is interesting to note that the "diffusion constant" κ which governs the spreading of the waves with time is proportional to the square of Planck's constant. Classical mechanics follows from quantum mechanics by letting Planck's constant go to zero: $\hbar \to 0$. It follows from (19.18) that in that limit the diffusion constant of the matter waves goes to zero. This reflects the fact that in classical mechanics the probability of the presence of a particle does not spread out with time.

Problem d Use the substitutions (19.18) to show that the Green's function for the Schrödinger equation in N dimensions is given by:

$$G(\mathbf{r}, t) = C\frac{1}{(2\pi i\hbar t/m)^{N/2}}e^{imr^2/2\hbar t}. \tag{19.19}$$

This Green's function plays a crucial role in the formulation of the *Feynman path integrals* which have been a breakthrough in quantum mechanics as well as in other fields. A clear description of the Feynman path integrals is given by Feynman and Hibbs [40].

Problem e Sketch the real part of $e^{imr^2/2\hbar t}$ in the Green's function for a fixed time as a function of radius r. Does the wavelength of the Green's function increase or decrease with distance?

The Green's function (19.19) actually has an interesting physical meaning which is based on the fact that it describes the propagation of matter waves injected at $t = 0$ at the origin. The Green's function can be written as $G = C\,(2\pi i\hbar t/m)^{-N/2}\, e^{i\Phi}$, where the phase of the Green's function is given by

$$\Phi = \frac{mr^2}{2\hbar t}. \tag{19.20}$$

As you noted in problem e of Section 8.4 the wavenumber of the waves depends on position. For a plane wave $e^{i\mathbf{k}\cdot\mathbf{r}}$ the phase is given by $\Phi = (\mathbf{k}\cdot\mathbf{r})$ and the wavenumber follows by taking the gradient of this function.

Problem f Show that for a plane wave

$$\mathbf{k} = \nabla\Phi. \tag{19.21}$$

Relation (19.21) has a wider applicability than plane waves. It has been shown by Whitham [119] that for a general phase function $\Phi(\mathbf{r})$ that varies smoothly with $\mathbf{r}$ the local wavenumber $\mathbf{k}(\mathbf{r})$ is defined by (19.21).

Problem g Use this to show that for the Green's function of the Schrödinger equation the local wavenumber is given by

$$\mathbf{k} = \frac{m\mathbf{r}}{\hbar t}. \tag{19.22}$$

Problem h Use the definition $\mathbf{v} = \mathbf{r}/t$ to show that this expression is equivalent to:

$$\mathbf{v} = \frac{\hbar\mathbf{k}}{m}. \tag{8.21}$$

In problem e you discovered that for a fixed time, the wavelength of the waves decreases when the distance r to the source is increased. This is consistent with (8.21): when a particle has moved further away from the source in a fixed time, its velocity is larger. This corresponds according to (8.21) with a larger wavenumber and hence with a smaller wavelength. This is indeed the behavior that is exhibited by the full wave-function (19.19).

The analysis in this chapter was not rigorous because the substitution $t \rightarrow (i/\hbar)t$ implies that the independent parameter is purely imaginary rather than real. This means that all the arguments used in the previous section for the complex integration should be carefully reexamined. However, a more rigorous analysis shows that (19.19) is indeed the correct Green's function for the Schrödinger equation [40]. The approach taken in this section shows that an educated guess can be very useful in deriving new results. One can in fact argue that many innovations in mathematical physics have been obtained using intuition or analogies rather than formal derivations. Of course, a formal derivation should ultimately substantiate results obtained from a more intuitive approach.

19.3 Helmholtz equation in one, two, and three dimensions

The Helmholtz equation plays an important role in mathematical physics because it is closely related to the wave equation. A complete analysis of the Green's function for the wave equation and the Helmholtz equation in different dimensions is given by DeSanto [34]. The Green's function for the wave equation for a medium with constant velocity c satisfies:

$$\nabla^2 G(\mathbf{r}, t; \mathbf{r}_0, t_0) - \frac{1}{c^2}\frac{\partial^2 G(\mathbf{r}, t; \mathbf{r}_0, t_0)}{\partial t^2} = \delta(\mathbf{r} - \mathbf{r}_0)\delta(t - t_0). \tag{19.23}$$

As shown in Section 19.1 the Green's function depends for a constant velocity c only on the relative location $\mathbf{r} - \mathbf{r}_0$ and the relative time $t - t_0$ so that without loss of generality we can take the source at the origin ($\mathbf{r}_0 = 0$) and let the source act at time $t_0 = 0$. In addition it follows from symmetry considerations that the Green's function depends only on the relative distance $|\mathbf{r} - \mathbf{r}_0|$ and not on the orientation of the vector $\mathbf{r} - \mathbf{r}_0$. This means that the Green's function then satisfies $G(\mathbf{r}, t; \mathbf{r}_0, t_0) = G(|\mathbf{r} - \mathbf{r}_0|, t - t_0)$ and we need to solve the following equation:

$$\nabla^2 G(r, t) - \frac{1}{c^2}\frac{\partial^2 G(r, t)}{\partial t^2} = \delta(\mathbf{r})\delta(t). \tag{19.24}$$

Problem a Under which conditions is this approach justified?

Problem b Use a similar treatment to that in Section 19.1 to show that when the Fourier transform (15.43) is used the Green's function satisfies the following equation in the frequency domain:

$$\nabla^2 G(r,\omega) + k^2 G(r,\omega) = \delta(\mathbf{r}), \tag{19.25}$$

where the wavenumber k satisfies $k = \omega/c$.

This equation is called the *Helmholtz equation* and is the reformulation of the wave equation in the frequency domain. In the following we suppress the factor ω in the Green's function but it should be remembered that the Green's function depends on frequency.

Let us first solve (19.25) for one dimension. In that case the Green's function is defined by

$$\frac{d^2 G}{dx^2} + k^2 G = \delta(x). \tag{19.26}$$

Problem c For $x \neq 0$ the right-hand side of this equation is equal to zero. Use this to show that for $x \neq 0$ the solution is given by

$$G(x) = A\ e^{ikx} + B\ e^{-ikx}, \tag{19.27}$$

where A and B are integration constants which need to be determined for the regions $x < 0$ and $x > 0$ separately.

In the Fourier transformation to the time domain, both terms are multiplied by $e^{-i\omega t}$. Using the relation $k = \omega/c$, this means that in the time domain the term e^{ikx} becomes $e^{-i\omega(t-x/c)}$. For increasing time t, the phase of the wave remains constant when x increases. This means that the term e^{ikx} corresponds in the time domain to a right-going wave. Similarly, the term e^{-ikx} describes a wave in the time domain that moves to the left. We are looking here for a Green's function that describes waves that move *away* from the source. This means that for $x > 0$ only the term e^{ikx} contributes, so in that region we must have $B = 0$, while for $x < 0$ only the term e^{-ikx} contributes hence we must take $A = 0$. This means that the Green's function is given by

$$G(x) = \begin{cases} B\ e^{-ikx} & \text{for} \quad x < 0 \\ A\ e^{ikx} & \text{for} \quad x > 0 \end{cases}. \tag{19.28}$$

The constants A and B follow from the requirements that $G(x)$ is continuous at $x = 0$ and that the first derivative is discontinuous at that point.

Problem d Use the theory of Section 18.3 to show that the jump in the first derivative at $x = 0$ is given by

$$\left[\frac{dG}{dx}\right]_{x=0-\varepsilon}^{x=0+\varepsilon} = 1. \tag{19.29}$$

Problem e Use these results to derive that $A = B = -i/2k$, and use this result to write the Green's function of the Helmholtz equation in one dimension as

$$G^{1D}(x) = \frac{-i}{2k} e^{ik|x|}. \tag{19.30}$$

Now we solve (19.25) for two and three space dimensions. To do this we consider the case of N dimensions, where N is either 2 or 3. Because the problem is spherically symmetric, we just need to consider a Green's function that depends on radius only.

Problem f Use (10.35) and (10.36) to show that for such a radially symmetric function in two or three dimensions the Laplacian is given by

$$\nabla^2 G(r) = \frac{1}{r^{N-1}} \frac{\partial}{\partial r}\left(r^{N-1} \frac{\partial G}{\partial r}\right). \tag{19.31}$$

The differential equation for the Green's function in N dimensions is thus given by

$$\frac{1}{r^{N-1}} \frac{\partial}{\partial r}\left(r^{N-1} \frac{\partial G}{\partial r}\right) + k^2 G(r, \omega) = \delta(\mathbf{r}). \tag{19.32}$$

This differential equation is not difficult to solve for two or three space dimensions for locations away from the source ($r \neq 0$). However, we need to consider carefully how the source $\delta(\mathbf{r})$ should be coupled to the solution of the differential equation. This can be achieved by integrating (19.25) over a sphere of radius R centered at the source and letting the radius go to zero.

Problem g Integrate (19.25) over this volume, use Gauss's law and let the radius R go to zero to show that the Green's function satisfies

$$\oint_{S_R} \frac{\partial G}{\partial r} dS = 1, \tag{19.33}$$

where the surface integral is over a sphere S_R with radius R in the limit $R \downarrow 0$. Show that this can also be written as

$$\lim_{r \downarrow 0} S_r \frac{\partial G}{\partial r} = 1, \tag{19.34}$$

where S_r is the surface of a sphere in N dimensions with radius r.

Note that the surface of the sphere in general goes to zero as $r \downarrow 0$ (except in one dimension), which implies that $\partial G/\partial r$ must be infinite in the limit $r \downarrow 0$ in more than one space dimension.

The differential equation (19.32) is a second order differential equation. Such an equation must be supplemented with two boundary conditions. The first boundary condition is given by (19.34), and specifies how the solution is coupled to the source at $\mathbf{r} = 0$. The second boundary condition that we will use reflects the fact that the waves generated by the source will move away from the source. The solutions that we will find will behave for large distances as $e^{\pm ikr}$, but it is not clear whether we should use the upper sign (+) or the lower sign (−).

Problem h Use the Fourier transform (15.42) and the relation $k = \omega/c$ to show that the integrand in the Fourier transformation to the time domain is proportional to $e^{-i\omega(t\mp r/c)}$.[‡] Show that the waves only move away from the source for the upper sign by showing that for this solution the distance r must increase for increasing time t to keep the phase constant. This means that this boundary condition dictates that the solution behaves in the limit $r \rightarrow \infty$ as e^{+ikr}.

The derivative of function e^{+ikr} is given by ike^{+ikr}, that is the derivative is ik times the original function. When the Green's function behaves for large r as e^{+ikr}, then the derivative of the Green's function must satisfy the same relation as the derivative of e^{+ikr}. This means that the Green's function satisfies for large distance r:

$$\frac{\partial G}{\partial r} = ikG. \tag{19.35}$$

This relation specifies that the energy radiates *away* from the source. For this reason (19.35) is called the *radiation boundary condition*.

Now we are at the point where we can actually construct the solution for each dimension. Before we go to two dimensions we first solve the Green's function in three dimensions.

Problem i Make for three dimensions ($N = 3$) the substitution $G(r) = f(r)/r$ and show that (19.32) implies that away from the source the function $f(r)$ satisfies

$$\frac{\partial^2 f}{\partial r^2} + k^2 f = 0. \tag{19.36}$$

This equation has the solution $Ce^{\pm ikr}$. According to problem h the upper sign should be used and the Green's function is given by $G(r) = Ce^{ikr}/r$. Show that condition (19.34) dictates that $C = -1/4\pi$, so that in three dimensions the Green's function is given by:

[‡] The convention of the notation $\mp$ is that the upper sign in $e^{-i\omega(t\mp r/c)}$ corresponds to the upper sign in $e^{\pm ikr}$. Similarly, the lower signs in both expressions correspond to each other. This means that the solution $e^{-i\omega(t-r/c)}$ corresponds to e^{+ikr} and that the solution $e^{-i\omega(t+r/c)}$ corresponds to e^{-ikr}.

$$G^{3D}(r) = \frac{-1}{4\pi}\frac{e^{ikr}}{r}. \tag{19.37}$$

The problem is actually most difficult in two dimensions because in that case the Green's function cannot be expressed in the simplest elementary functions.

Problem j Show that in two dimensions ($N = 2$) the differential equation of the Green's function away from the source is given by

$$\frac{\partial^2 G}{\partial r^2} + \frac{1}{r}\frac{\partial G}{\partial r} + k^2 G(r) = 0, \qquad r \neq 0. \tag{19.38}$$

Problem k This equation cannot be solved in terms of elementary functions. However, there is a close relation between (19.38) and the Bessel equation that is given by

$$\frac{d^2 F}{dx^2} + \frac{1}{x}\frac{dF}{dx} + \left(1 - \frac{m^2}{x^2}\right) F = 0. \tag{19.39}$$

Show that the $G(kr)$ satisfies the Bessel equation for order $m = 0$.

This implies that the Green's function is given by the solution of the zeroth order Bessel equation with argument kr. The Bessel equation is a second order differential equation, therefore there are two independent solutions. The solution that is finite everywhere is denoted by $J_m(x)$ and is called the regular Bessel function. The second solution is singular at the point $x = 0$ and is called the Neumann function and denoted by $N_m(x)$. The Green's function is obviously a linear combination of $J_0(kr)$ and $N_0(kr)$. In order to determine how this linear combination is constructed it is crucial to consider the behavior of these functions at the source (i.e. for $kr = 0$) and at infinity (i.e. for $kr \gg 1$). The required asymptotic behavior can be found in textbooks such as Butkov [24] and Arfken [5] and is summarized in Table 19.1.

Problem l Show that neither $J_0(kr)$ nor $N_0(kr)$ behaves for large values of r as e^{+ikr}. Show that the linear combination $J_0(kr) + i N_0(kr)$ does behave as e^{+ikr}.

The Green's function thus is a linear combination of the regular Bessel function and the Neumann function. This particular combination is called the *first Hankel function of degree zero* and is denoted by $H_0^{(1)}(kr)$. The Hankel functions are simply

Table 19.1 *Leading asymptotic behavior of the Bessel function and Neumann function of order zero*

	$J_0(x)$	$N_0(x)$
$x \to 0$	$1 - \frac{1}{4}x^2 + O(x^4)$	$\frac{2}{\pi}\ln(x) + O(1)$
$x \gg 1$	$\sqrt{\frac{2}{\pi x}}\cos\left(x - \frac{\pi}{4}\right) + O(x^{-3/2})$	$\sqrt{\frac{2}{\pi x}}\sin\left(x - \frac{\pi}{4}\right) + O(x^{-3/2})$

linear combinations of the Bessel function and the Neumann function:

$$\left.\begin{aligned} H_m^{(1)}(x) &\equiv J_m(x) + iN_m(x), \\ H_m^{(2)}(x) &\equiv J_m(x) - iN_m(x). \end{aligned}\right\} \tag{19.40}$$

Problem m Use this definition and Table 19.1 to show that $H_0^{(1)}(kr)$ behaves for large values of r as $e^{+ikr-i\pi/4}/\sqrt{(\pi/2)kr}$ and that in this limit $H_0^{(2)}(kr)$ behaves as $e^{-ikr-i\pi/4}/\sqrt{(\pi/2)kr}$. Use this to argue that the Green's function is given by

$$G(r) = CH_0^{(1)}(kr), \tag{19.41}$$

where the constant C still needs to be determined.

Problem n This constant follows from the requirement (19.34) at the source. Use (19.40) and the asymptotic value of the Bessel function and the Neumann function given in Table 19.1 to derive the asymptotic behavior of the Green's function near the source and use this to show that $C = -i/4$.

This result implies that in two dimensions the Green's function of the Helmholtz equation is given by

$$G^{2D}(r) = \frac{-i}{4}H_0^{(1)}(kr). \tag{19.42}$$

Summarizing these results and reverting to the more general case of a source at location $\mathbf{r}_0$ it follows that in one, two, and three dimensions the Green's functions of the Helmholtz equation are given by

$$\left.\begin{aligned} G^{1D}(x, x_0) &= \frac{-i}{2k} e^{ik|x-x_0|}, \\ G^{2D}(\mathbf{r}, \mathbf{r}_0) &= \frac{-i}{4} H_0^{(1)}(k\,|\mathbf{r}-\mathbf{r}_0|), \\ G^{3D}(\mathbf{r}, \mathbf{r}_0) &= \frac{-1}{4\pi} \frac{e^{ik|\mathbf{r}-\mathbf{r}_0|}}{|\mathbf{r}-\mathbf{r}_0|}. \end{aligned}\right\} \tag{19.43}$$

Note that in two and three dimensions the Green's function is singular at the source $\mathbf{r}_0$.

Problem o Show that these singularities are *integrable*, that is show that when the Green's function is integrated over a sphere with finite radius around the source the result is finite.

There is a physical reason why the Green's function in two and three dimensions has an integrable singularity. Suppose one has a source that is not a point source but that the source is constant within a sphere with radius R centered around the origin. The response p to this source is given by $p(\mathbf{r}) = \int_{r'<R} G(\mathbf{r}, \mathbf{r}')dV'$, where the integration over the variable $\mathbf{r}'$ is over a sphere with radius R. It follows from this expression that the response at the origin is given by

$$p(\mathbf{r}=0) = \int_{r'<R} G(\mathbf{r}=0, \mathbf{r}')\,dV'. \tag{19.44}$$

Since the excitation of this field is finite everywhere, the response $p(\mathbf{r} = 0)$ should be finite. This implies that the integral (19.44) should be finite as well, which is a different way of stating that the singularity of the Green's function must be integrable.

19.4 Wave equation in one, two, and three dimensions

In this section we consider the Green's function for the wave equation in one, two, and three dimensions. This means that we consider solutions to the wave equation with an impulsive source at location $\mathbf{r}_0$ at time t_0:

$$\nabla^2 G(\mathbf{r}, t; \mathbf{r}_0, t_0) - \frac{1}{c^2}\frac{\partial^2 G(\mathbf{r}, t; \mathbf{r}_0, t_0)}{\partial t^2} = \delta(\mathbf{r}-\mathbf{r}_0)\delta(t-t_0). \tag{19.23}$$

It was shown in the previous section that this Green's function depends only on the relative distance $|\mathbf{r}-\mathbf{r}_0|$ and the relative time $t - t_0$. For the case of a source at the origin ($\mathbf{r}_0 = 0$) acting at time zero ($t_0 = 0$) the time domain solution follows by applying a Fourier transform to the solution $G(\mathbf{r}, \omega)$ of the previous section. This Fourier transform is simplest in three dimensions, hence we will start with this case.

Problem a Apply the Fourier transform (see (15.42)) to the three-dimensional Green's function (19.37) and use the relation $k = \omega/c$ and the properties of the delta function to show that the Green's function is given in the time domain by

$$G^{3D}(\mathbf{r}, t) = -\frac{1}{4\pi r}\delta\left(t - \frac{r}{c}\right). \tag{19.45}$$

Problem b Consider the wave equation with a general source term $S(\mathbf{r},t)$:

$$\nabla^2 p(\mathbf{r}, t) - \frac{1}{c^2}\frac{\partial^2 p(\mathbf{r}, t_0)}{\partial t^2} = S(\mathbf{r}, t). \tag{19.46}$$

Use the Green's function (19.45) to show that a solution of this equation is given by

$$p(\mathbf{r}, t) = -\frac{1}{4\pi}\int \frac{S\left(\mathbf{r}', t - \frac{|\mathbf{r} - \mathbf{r}'|}{c}\right)}{|\mathbf{r} - \mathbf{r}'|} dV'. \tag{19.47}$$

Note that since $\left|\mathbf{r} - \mathbf{r}'\right|$ is always positive, the response $p(\mathbf{r}, t)$ depends only on the source function at *earlier* times. The solution therefore has a causal behavior and the Green's function (19.45) is called the *retarded Green's function*. However, in several applications one does not want to use a Green's function that depends on excitation at earlier times. An example is reflection seismology, in which one records the wave field at the surface, and from these observations one wants to reconstruct the wave field at *earlier* times while it was being reflected off layers inside the Earth. (See the treatment in Section 8.3 and the work of Schneider [93].) A Green's function with waves that propagate towards the source and are then annihilated by the source can be obtained by replacing the radiation condition (19.35) by $\partial G/\partial r = -ikG$. The only difference is the minus sign on the right-hand side, which is equivalent to replacing k by $-k$.

Problem c Apply the Fourier transform (15.42) to the three-dimensional Green's function (19.37) with k replaced by $-k$ and show that the resulting Green's function in the time is given by

$$G^{3D,\ advanced}(\mathbf{r}, t) = -\frac{1}{4\pi r}\delta\left(t + \frac{r}{c}\right), \tag{19.48}$$

and that the following function is a solution of the wave equation (19.46):

$$p(\mathbf{r},t) = -\frac{1}{4\pi}\int \frac{S\left(t + \frac{|\mathbf{r}-\mathbf{r}'|}{c}\right)}{|\mathbf{r}-\mathbf{r}'|}dV'. \tag{19.49}$$

Note that in this representation the wave field is expressed in terms of the source function at *later* times. For this reason the Green's function (19.48) is called the *advanced Green's function*. The fact that the wave equation has both a retarded and an advanced solution is mathematically due to the wave equation (19.46) being invariant for time-reversal. This means that when one replaces t by $-t$ the equation does not change. Physically this means that the wave equation does not know the "direction of time". In practice one most often works with the retarded Green's function, but keep in mind that in some applications, such as exploration seismology, the advanced Green's functions are crucial. In the remaining part of this section we will focus exclusively on the retarded Green's functions that represent causal solutions.

In order to obtain the Green's function for two dimensions in the time domain one could apply a Fourier transform to the solution (19.42). This involves taking the Fourier transform of a Hankel function, and it is not obvious how this Fourier integral should be solved (although it can be solved). Here we follow an alternative route by recognizing that the Green's function in two dimensions is identical to the solution of the wave equation in three dimensions when the source is not a point source but a cylindrical source. In other words, we obtain the two-dimensional Green's function by considering the wave field in three dimensions that is generated by a source which is distributed homogeneously along the z-axis. In order to separate the distance to the origin from the distance to the z-axis the variables r and ρ are used, see Figure 19.2.

Problem d Use this line source to show that

$$G^{2D}(\rho,t) = \int_{-\infty}^{\infty} G^{3D}(r,t)\,dz. \tag{19.50}$$

Problem e Use the Green's function (19.45) and the relation

$$r = \sqrt{\rho^2 + z^2}$$

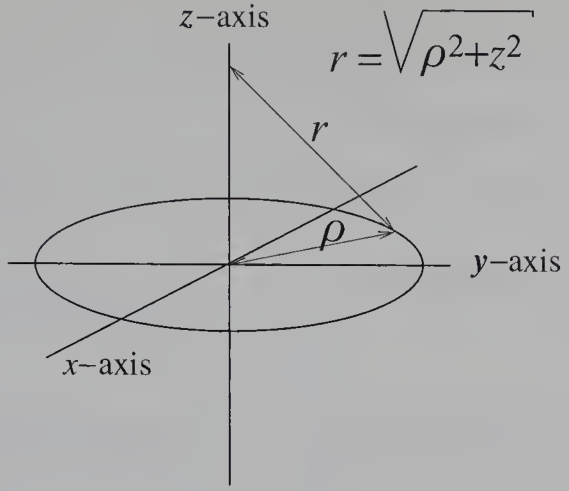

Fig. 19.2 Definition of the variables r and ρ.

to show that

$$G^{2D}(\rho, t) = -\frac{1}{2\pi}\int_0^\infty \frac{\delta\left(t - \frac{\sqrt{\rho^2+z^2}}{c}\right)}{\sqrt{\rho^2+z^2}}\,dz. \tag{19.51}$$

Note that the integration interval has been changed from $(-\infty, \infty)$ to $(0, \infty)$, and show how this can be achieved.

The distance r in three dimensions no longer appears in this expression.

Problem f The integral (19.51) can be solved by introducing the new integration variable $u \equiv \sqrt{\rho^2 + z^2}$, instead of the old integration variable z. Show that with this new variable the integral (19.51) can be written as

$$G^{2D}(\rho, t) = -\frac{1}{2\pi}\int_\rho^\infty \frac{\delta\left(t - \frac{u}{c}\right)}{\sqrt{u^2-\rho^2}}du: \tag{19.52}$$

pay attention to the limits of integration!

Problem g Use the property $\delta(ax) = \delta(x)/|a|$ to rewrite this integral and evaluate the resulting integral separately for $t < \rho/c$ and $t > \rho/c$. Finally denote the distance ρ in the two-dimensional (x, y)-plane by r to show that

$$G^{2D}(r, t) = \begin{cases} 0 & \text{for} \quad t < r/c \\ -\frac{1}{2\pi}\frac{1}{\sqrt{t^2 - r^2/c^2}} & \text{for} \quad t > r/c \end{cases}. \tag{19.53}$$

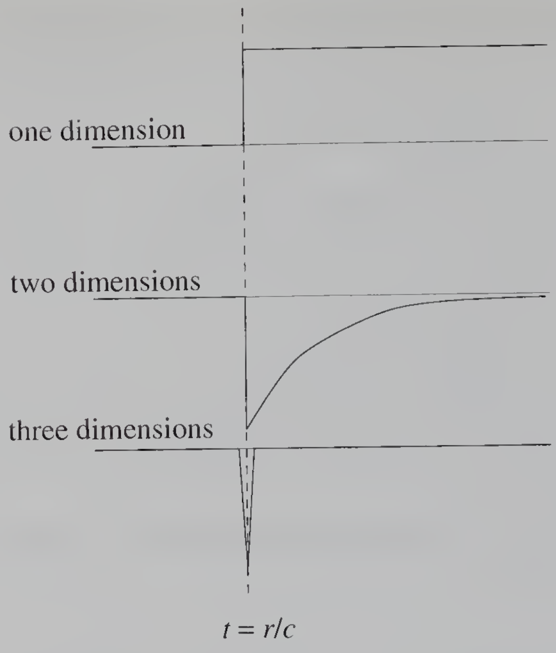

Fig. 19.3 The Green's function of the wave equation in one, two, and three dimensions as a function of time.

This Green's function and the Green's function for the three-dimensional case are sketched in Figure 19.3. There is a fundamental difference between the Green's function for two dimensions and the Green's function (19.45) for three dimensions. In three dimensions the Green's function is a delta-function $\delta(t - r/c)$ modulated by the geometrical spreading $-1/4\pi r$. This means that the response to a delta-function source has the same shape as the input function $\delta(t)$ that excites the wave field. An impulsive input leads to an impulsive output with a time delay given by r/c and the solution is only nonzero at the wave front $t = r/c$. However, (19.53) shows that an impulsive input in two dimensions leads to a response that is not impulsive. The response has an infinite duration and decays with time as $1/\sqrt{t^2 - r^2/c^2}$: the solution is not only nonzero *at* the wave front $t = r/c$, it is nonzero everywhere *within* this wave front. This means that in two dimensions an impulsive input leads to a sound response that is of infinite duration. Following my teacher Professor Eckhaus, one can therefore say that:

> *Any word spoken in two dimensions will reverberate forever (albeit weakly).*

The approach we have taken to compute the Green's function in two dimensions is interesting in that we solved the problem first in a higher dimension and retrieved the solution by integrating over one space dimension. Note that for this trick it is not necessary that this higher-dimensional space indeed exists! (Although in this case it does.) Remember that we took this approach because we did not want to evaluate the Fourier transform of a Hankel function. We can also turn this around: the Green's function (19.53) can be used to determine the Fourier transform of the Hankel function.

Problem h Show that the Fourier transform of the Hankel function is given by:

$$\int_{-\infty}^{\infty} H_0^{(1)}(x)e^{-iqx}dx = \begin{cases} 0 & \text{for} \quad q < 1 \\ \dfrac{2}{i\pi}\dfrac{1}{\sqrt{q^2-1}} & \text{for} \quad q > 1 \end{cases} . \tag{19.54}$$

Hint: take the Fourier transform $G^{2D}(\mathbf{r}, \mathbf{r}_0)$ in (19.43) in order to obtain the Green's function for two dimensions in the time domain and compare the result with the corresponding expression (19.53). Make a suitable change of variables to arrive at (19.54).

Let us continue with the Green's function of the wave equation in one dimension in the time domain.

Problem i Use the Green's function for one dimension of the last section to show that in the time domain

$$G^{1D}(x,t) = -\frac{ic}{4\pi}\int_{-\infty}^{\infty}\frac{1}{\omega}e^{-i\omega(t-|x|/c)}d\omega. \tag{19.55}$$

This integral resembles the integral used for the calculation of the Green's function in three dimensions. The only difference is the term $1/\omega$ in the integrand, and because of this term we cannot immediately evaluate the integral. However, the $1/\omega$ term can be removed by differentiating (19.55) with respect to time, and the remaining integral can then be evaluated analytically.

Problem j Show that

$$\frac{\partial G^{1D}(x,t)}{\partial t} = \frac{c}{2}\delta\left(t - \frac{|x|}{c}\right). \tag{19.56}$$

Problem k This expression can be integrated but one condition is needed to specify the integration constant that appears. We will use here that at $t = -\infty$ the Green's function vanishes. Show that with this condition the Green's function is given by:

$$G^{1D}(x,t) = \begin{cases} 0 & \text{for} \quad t < |x|/c \\ c/2 & \text{for} \quad t > |x|/c \end{cases} . \tag{19.57}$$

Just as in two dimensions the solution is nonzero everywhere *within* the expanding wave front and not only on the wave front $|x| = ct$ as in three dimensions. However, there is an important difference: in two dimensions the solution varies for all moments with time whereas in one dimension the solution is constant except for $t = |x|/c$. The human ear is only sensitive to pressure variations, it is insensitive to a static pressure. Therefore a one-dimensional human will only hear a sound at $t = |x|/c$ but not at later times.

In order to appreciate the difference in sound propagation in one, two, and three space dimensions, the Green's functions for the different dimensions are sketched in Figure 19.3. Note the dramatic change in the response for different numbers of dimensions. This change in the properties of the Green's function with change in dimension has been used somewhat jokingly by Morley [71] to give "a simple proof that the world is three dimensional." When you worked through Sections 19.1 and 19.2 you learned that for both the heat equation and the Schrödinger equation the solution does not depend fundamentally on the number of dimensions. This is in stark contrast with the solutions of the wave equation which depend critically on the number of dimensions.

19.5 If I can hear you, you can hear me

In this section we treat a property of the Green's function of acoustic waves. We start with the acoustic wave equation:

$$\nabla \cdot \left(\frac{1}{\rho(\mathbf{r})} \nabla p \right) + \frac{\omega^2}{\kappa(\mathbf{r})} p = f \, . \tag{8.7}$$

The density $\rho(\mathbf{r})$ and the bulk modulus $\kappa(\mathbf{r})$ can be arbitrary functions of position. Let a solution p_1 be excited by an excitation f_1 and a solution p_2 by an excitation f_2. As shown in Section 8.3 the two solutions are related by the following expression

$$\oint_S \frac{1}{\rho} \left(p_2 \nabla p_1 - p_1 \nabla p_2 \right) \cdot d\mathbf{S} = \int_V \{ p_2 f_1 - p_1 f_2 \} \, dV \, . \tag{8.11}$$

Let us consider a volume V that is either bounded by a free surface where $p = 0$, or that extends to infinity. The parts of the boundary that form a free surface do not contribute to the surface integral in the left-hand side, because at the free surface $p = 0$, so that the integrand vanishes.

Problem a At the parts of the surface S that are placed at infinity, the radiation boundary condition (19.35) applies. Show that with this boundary condition

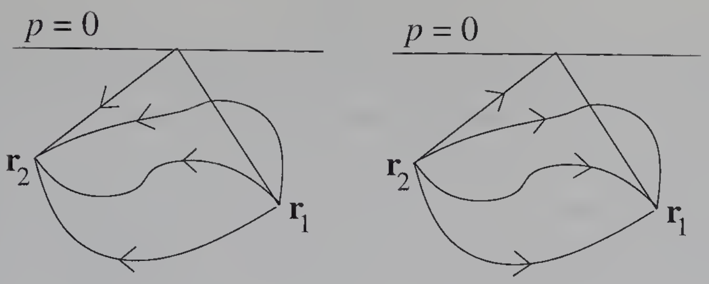

Fig. 19.4 Paths that connect the points $\mathbf{r}_1$ and $\mathbf{r}_2$ in opposite directions. The waves that propagate in the two opposite directions are identical.

the contributions of the surface S to the integral in the left-hand side of (8.11) vanishes.

This means that for the volume under consideration the left-hand side of (8.11) vanishes. This result holds for a general excitation $f_{1,2}$. Let us take $f_1(\mathbf{r}) = \delta(\mathbf{r} - \mathbf{r}_1)$, a point source at location $\mathbf{r}_1$. The response to this excitation is the Green's function: $p_1(\mathbf{r}) = G(\mathbf{r}, \mathbf{r}_1)$. Similarly, we take for the excitation of p_2 a delta function placed at location $\mathbf{r}_2$, so that $f_2(\mathbf{r}) = \delta(\mathbf{r} - \mathbf{r}_2)$. The corresponding pressure field is given by the Green's function $p_2(\mathbf{r}) = G(\mathbf{r}, \mathbf{r}_2)$.

Problem b Insert these solutions into equation (8.11), use that the left-hand side vanishes, and show that

$$G(\mathbf{r}_1, \mathbf{r}_2) = G(\mathbf{r}_2, \mathbf{r}_1). \tag{19.58}$$

This result is called the *reciprocity theorem*. The Green's function $G(\mathbf{r}_1, \mathbf{r}_2)$ gives the pressure at location $\mathbf{r}_1$ due to a point source at $\mathbf{r}_2$, while the Green's function $G(\mathbf{r}_2, \mathbf{r}_1)$ gives the pressure at location $\mathbf{r}_2$ due to a point source at $\mathbf{r}_1$. According to expression (19.58) these solutions are identical. This means that the role of the source and receiver can be interchanged, and the recorded pressure field is identical. This situation is sketched in Figure 19.4. Note that the density and bulk modulus can be arbitrary functions of space. This means that the waves may travel along a multitude of paths. Since we allowed for the presence of a free surface, the pressure waves may also bounce off the free surface. Intuitively, it is clear that the travel time from $\mathbf{r}_1$ to $\mathbf{r}_2$ is the same as the travel time from $\mathbf{r}_2$ to $\mathbf{r}_1$. However, the reciprocity theorem (19.58) states that this holds for the amplitude as well, and that the complete waveforms are identical.

This means that if I can hear you, you can hear me, regardless of how complex the medium is, and how many many echoes it may produce. (Provided that our sense of hearing is equally good.) This also implies that if one submarine can detect another submarine with its sonar, then the second submarine can detect the first one as well. (Assuming that their sonar equipment is equally sensitive.) This property holds no matter how complex the distribution of the speed of sound in the ocean may be. The reciprocity theorem does not hold only for acoustic waves. Elastic waves satisfy a similar reciprocity theorem that incorporates the vector character of elastic waves [3, 43]. Reciprocity holds as well for the solutions of the heat equation (19.1) no matter what the spatial distribution of the conductivity $\kappa(\mathbf{r})$ is.

20

Normal modes

Many physical systems have the property that they can carry out oscillations at certain specific frequencies only. As a child (and hopefully also as an adult) you will have discovered that a swing in a playground will move only with a specific natural period, and that the force that pushes the swing is only effective when the period of the force matches the period of the swing. The patterns of motion at which a system oscillates are called the *normal modes* of the system. A swing has one normal mode, but you have seen in Section 13.6 that a simple model of a tri-atomic molecule has three normal modes. An example of a normal mode of a system is given in Figure 20.1 which shows the pattern of oscillation of a metal plate which is driven by an oscillator at a fixed frequency. The screw in the middle of the plate shows the point at which the force on the plate is applied. Sand is sprinkled on the plate. When the frequency of the external force is equal to the frequency of a normal mode of the plate, the motion of the plate is given by the motion that corresponds to that specific normal mode. Such a pattern of oscillation has nodal lines where the motion vanishes. These nodal lines are visible because the sand on the plate collects at these lines.

In this chapter, the normal modes of a variety of systems are analyzed. Normal modes play an important role in a variety of applications because the eigenfrequencies of normal modes provide important information of physical systems. Examples include the spectral lines of light emitted by atoms which have led to the advent of quantum mechanics and its description of the structure of atoms, and the normal modes of the Earth which provide information about the structure of our planet. In addition, normal modes are used in this chapter to introduce some properties of special functions, such as Bessel functions and Legendre functions. This is achieved by analyzing the normal modes of a system in one, two, and three dimensions in Sections 20.1–20.3.

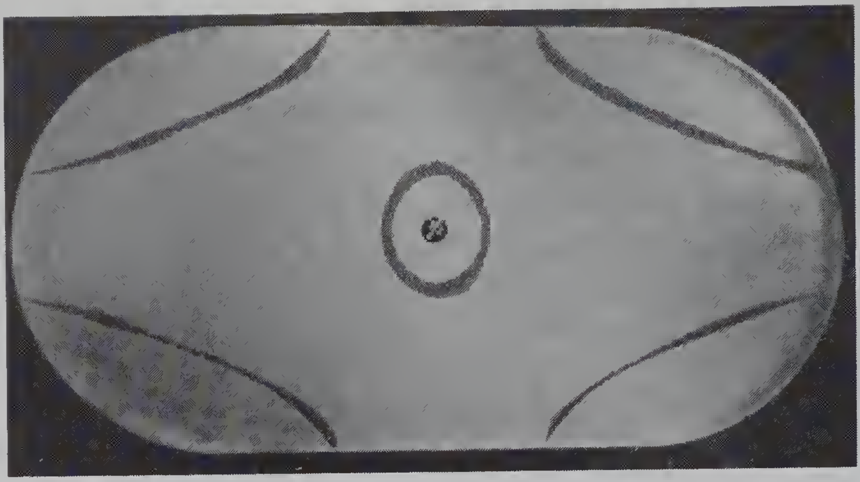

Fig. 20.1 Sand on a metal plate which is driven by an oscillator at a frequency that corresponds to one of the eigenfrequencies of the plate. This figure was prepared by John Scales at the Physical Acoustics Laboratory at the Colorado School of Mines.

20.1 Normal modes of a string

In this and the following two sections we assume that the motion of the system is governed by the Helmholtz equation

$$\nabla^2 u + k^2 u = 0. \tag{20.1}$$

In this expression the wavenumber k is related to the angular frequency ω by the relation

$$k = \frac{\omega}{c}. \tag{20.2}$$

For simplicity we assume the system to be homogeneous, which means that the velocity c is constant. This in turn implies that the wavenumber k is constant. In Sections 20.1–20.3 we consider a body of radius R. Since a circle or a sphere of radius R has a diameter $2R$ we will consider here a string of length $2R$ in order to be able to make meaningful comparisons. We assume that the endpoints of the string are fixed so that the boundary conditions are:

$$u(0) = u(2R) = 0. \tag{20.3}$$

Problem a Show that the solutions of (20.1) that satisfy the boundary conditions (20.3) are given by $\sin(k_n x)$ with the wavenumber k_n given by

$$k_n = \frac{n\pi}{2R}, \tag{20.4}$$

where n is an integer.

For a number of purposes it is useful to normalize the modes: this means that we require that the modes $u_n(x)$ satisfy the condition $\int_0^{2R} u_n^2(x)dx = 1$.

Problem b Show that the normalized modes are given by

$$u_n(x) = \frac{1}{\sqrt{R}} \sin(k_n x). \tag{20.5}$$

Problem c Sketch the modes for several values of n as a function of the distance x.

Problem d The modes $u_n(x)$ are orthogonal, which means that the inner product $\int_0^{2R} u_n(x)u_m(x)dx$ vanishes when $n \neq m$. Give a proof of this property by deriving that

$$\int_0^{2R} u_n(x)u_m(x)dx = \delta_{nm}. \tag{20.6}$$

We conclude from this section that the modes of a string are oscillatory functions with a wavenumber that can only have discrete well-defined values k_n. According to (20.2) this means that the string can only vibrate at discrete frequencies which are given by

$$\omega_n = \frac{n\pi c}{2R}. \tag{20.7}$$

This property may be familiar to you because you probably know that a guitar string vibrates only at very specific frequencies which determine the pitch of the sound that you hear. The results of this section imply that each string oscillates not only at one particular frequency, but at many discrete frequencies. The oscillation with the lowest frequency is given by (20.7) with $n = 1$: this is called the *fundamental mode* or *ground-tone*. This is what the ear perceives as the pitch of the tone. The oscillations corresponding to larger values of n are called the *higher modes* or *overtones*. The particular mix of overtones determines the timbre of the signal. If the higher modes are strongly excited the ear perceives this sound as metallic, whereas just the fundamental mode is perceived as a smooth sound. The reader who is interested in the theory of musical instruments can consult reference [88].

The discrete modes are not a peculiarity of the string. Most systems that support waves and that are of a finite extent support modes. For example, in Figure 15.1 the spectrum of the sound of a soprano saxophone is shown. This spectrum is characterized by well-defined peaks that correspond to the modes of the air-waves in the instrument. Mechanical systems in general have discrete modes, these modes can be destructive when they are excited at their resonance frequency. The matter waves in atoms are organized in modes as well, and this is ultimately the reason why atoms in an excited state emit light only at very specific frequencies, called spectral lines.

20.2 Normal modes of a drum

In the previous section we looked at the modes of a one-dimensional system. Here we derive the modes of a two-dimensional system which is a model of a drum. We consider a two-dimensional membrane that satisfies the Helmholtz equation (20.1). The membrane is circular and has radius R. At the edge $r = R$ the membrane cannot move, this means that in cylindrical coordinates the boundary condition for the waves $u(r, \varphi)$ is given by:

$$u(R, \varphi) = 0. \tag{20.8}$$

In order to find the modes of the drum we will use *separation of variables*, which means that we seek solutions that can be written as a product of a function that depends only on r and a function that depends only on φ:

$$u(r, \varphi) = F(r)G(\varphi). \tag{20.9}$$

Problem a Insert this solution into the Helmholtz equation, use the expression of the Laplacian in cylindrical coordinates, and show that the resulting equation can be written as

$$\underbrace{\left[\frac{1}{F(r)} r \frac{\partial}{\partial r}\left(r \frac{\partial F}{\partial r}\right) + k^2 r^2\right]}_{\text{(A)}} = \underbrace{-\frac{1}{G(\varphi)} \frac{\partial^2 G}{\partial \varphi^2}}_{\text{(B)}} . \tag{20.10}$$

Problem b The terms labeled (A) depend only on the variable r whereas the terms labeled (B) depend only on the variable φ. These terms can only be equal for all values of r and φ when they depend neither on r nor on φ, that is when they are a constant. Use this to show that $F(r)$ and $G(\varphi)$ satisfy the following differential equations:

$$\frac{d^2 F}{dr^2} + \frac{1}{r} \frac{dF}{dr} + \left(k^2 - \frac{\mu}{r^2}\right) F = 0, \tag{20.11}$$

$$\frac{d^2 G}{d\varphi^2} + \mu G = 0, \tag{20.12}$$

where μ is a constant that is not yet known.

These differential equations need to be supplemented with boundary conditions. The boundary conditions for $F(r)$ follow from the requirements that this function is finite everywhere and that the displacement vanishes at the edge of the drum:

$$F(r) \text{ is finite everywhere}, \qquad F(R) = 0. \tag{20.13}$$

The boundary condition for $G(\varphi)$ follows from the requirement that if we rotate the drum through 360°, every point on the drum returns to its original position. This means that the modes satisfy the requirement that $u(r,\varphi) = u(r,\varphi + 2\pi)$. This implies that $G(\varphi)$ satisfies the *periodic boundary condition*:

$$G(\varphi) = G(\varphi + 2\pi). \tag{20.14}$$

Problem c The general solution of (20.12) is given by

$$G(\varphi) = e^{\pm i\sqrt{\mu}\varphi}.$$

Show that the boundary condition (20.14) implies that $\mu = m^2$, with m an integer.

This means that the dependence of the modes on the angle φ is given by:

$$G(\varphi) = e^{im\varphi}. \tag{20.15}$$

The value $\mu = m^2$ can be inserted in (20.11). The resulting equation then bears a close resemblance to the Bessel equation:

$$\frac{d^2 J_m}{dx^2} + \frac{1}{x}\frac{dJ_m}{dx} + \left(1 - \frac{m^2}{x^2}\right) J_m = 0. \tag{20.16}$$

This equation has two independent solutions: the Bessel function $J_m(x)$ which is finite everywhere and the Neumann function $N_m(x)$ which is singular at $x = 0$. The Bessel functions $J_m(x)$ for several orders m are shown in Figure 20.2. All the Bessel functions are oscillatory functions which decay with increasing values of the argument x. Note that there is a phase shift of a quarter cycle between the successive

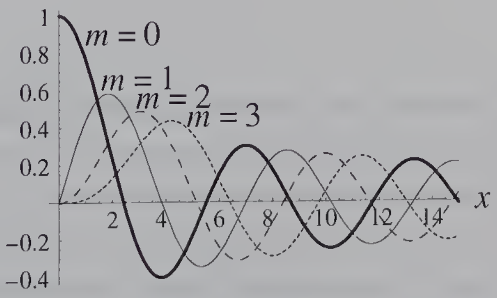

Fig. 20.2 The Bessel functions $J_m(x)$ for several orders m.

orders m. Near $x = 0$ the lowest order Bessel functions behave as $J_0(x) \sim 1$, while the higher order Bessel functions behave as $J_1(x) \sim x$ and $J_2(x) \sim x^2$. This is due to the fact that the Bessel function $J_m(x)$ behaves as x^m for small values of x.

Problem d Show that the general solution of (20.11) can be written as:

$$F(r) = AJ_m(kr) + BN_m(kr), \tag{20.17}$$

with A and B integration constants.

Problem e Use the boundary conditions of $F(r)$ to show that $B = 0$ and that the wavenumber k must take a value such that $J_m(kR) = 0$.

This last condition for the wavenumber is analogous to the condition (20.4) for the one-dimensional string. For both the string and the drum the wavenumber can only take discrete values: these values are dictated by the condition that the displacement vanishes at the outer boundary of the string or the drum. For the drum there are for every value of the angular degree m infinitely many wavenumbers that satisfy the requirement $J_m(kR) = 0$. These wavenumbers are labeled with a subscript n, but since these wavenumbers are different for each value of the angular order m, the allowed wavenumbers carry two indices and are denoted here by $k_n^{(m)}$. They satisfy the condition

$$J_m(k_n^{(m)}R) = 0. \tag{20.18}$$

The zeroes (or roots) of the Bessel function $J_m(x)$ are not known in closed analytical form. However, numerical tables exists for the roots of Bessel functions, see for example, Table 9.4 of Abramowitz and Stegun [1]. Take a look at this reference which contains a bewildering collection of formulas, graphs, and tables of mathematical functions. The lowest order zeroes of the Bessel functions $J_0(x), J_1(x), \ldots, J_5(x)$ are shown in Table 20.1.

Using the results in this section it follows that the modes of the drum are given by

$$u_{nm}(r, \varphi) = J_m(k_n^{(m)}r)e^{im\varphi}. \tag{20.19}$$

Problem f Let us first consider the φ-dependence of these modes. Show that when one follows the mode $u_{nm}(r, \varphi)$ along a complete circle around the origin one encounters exactly m oscillations of that mode.

Problem g Find the eigenfrequencies of the five modes of the drum with the lowest frequencies and make a sketch of the associated standing waves of the

Table 20.1 *The lowest roots of the Bessel function $J_m(x)$, these are the values of x for which $J_m(x) = 0$.*

n	$m = 0$	$m = 1$	$m = 2$	$m = 3$	$m = 4$	$m = 5$
1	2.40482	3.83171	5.13562	6.38016	7.58834	8.77148
2	5.52007	7.01559	8.41724	9.76102	11.06471	12.33860
3	8.65372	10.17347	11.61984	13.01520	14.37254	15.70017
4	11.79153	13.32369	14.79595	16.22347	17.61597	18.98013
5	14.93091	16.47063	17.95982	19.4092	20.82693	22.21780
6	18.07106	19.61586	21.11700	22.58273	24.01902	25.43034
7	21.21163	22.76008	24.27011	25.74817	27.19909	28.62662

drum. Use (20.18) and Table 20.1 to determine what the values of n and m are for these five modes. Figure 20.2 gives the dependence of the lowest order Bessel functions on their argument.

Problem h Use Table 20.1 to compute the separation between the different zero-crossings for a fixed value of m. To which number does this separation converge for the zero-crossings at large values of x?

The shape of the Bessel function is more difficult to see than the properties of the functions $e^{im\varphi}$. As shown in Section 9.7 of Butkov [24], these functions satisfy a large number of properties which include recursion relations and series expansions. However, at this point the following facts are most important:

- The Bessel functions $J_m(x)$ are oscillatory functions which decay with distance: in a sense they behave as decaying standing waves. We will return to this issue in Section 20.5.
- The Bessel functions satisfy an orthogonality relation similar to the orthogonality relation (20.6) for the modes of the string. This orthogonality relation is treated in more detail in Section 20.4.

20.3 Normal modes of a sphere

In this section we consider the normal modes of a spherical surface with radius R. We only consider the modes that are associated with the waves that propagate along the surface, hence we do not consider wave motion in the *interior* of the sphere. The modes are assumed to satisfy the wave equation (20.1). Since the waves propagate on the spherical surface, they are only a function of the angles θ and φ that are used in spherical coordinates: $u = u(\theta, \varphi)$. Using the Laplacian expressed in spherical coordinates, wave equation (20.1) is then given by

$$\frac{1}{R^2}\left[\frac{1}{\sin\theta}\frac{\partial}{\partial\theta}\left(\sin\theta\frac{\partial u}{\partial\theta}\right)+\frac{1}{\sin^2\theta}\frac{\partial^2 u}{\partial\varphi^2}\right]+k^2u=0. \tag{20.20}$$

Again, we seek a solution by applying separation of variables by writing the solution in a form similar to (20.9):

$$u(\theta,\varphi)=F(\theta)G(\varphi). \tag{20.21}$$

Problem a Insert this into (20.20) and apply separation of variables to show that $F(\theta)$ satisfies the following differential equation:

$$\sin\theta\frac{d}{d\theta}\left(\sin\theta\frac{dF}{d\theta}\right)+\left(k^2R^2\sin^2\theta-\mu\right)F=0, \tag{20.22}$$

and that $G(\varphi)$ satisfies (20.12), where the unknown constant μ does not depend on θ or φ.

To make further progress we have to apply boundary conditions. Just as with the drum in Section 20.2, the system is invariant when a rotation over 2π is applied: $u(\theta,\varphi)=u(\theta,\varphi+2\pi)$. This means that $G(\varphi)$ satisfies the same differential equation (20.12) as for the case of the drum and the same periodic boundary condition (20.14). The solution is therefore given by $G(\varphi)=e^{im\varphi}$ and the separation constant satisfies $\mu=m^2$, with m an integer. Using this, the differential equation for $F(\theta)$ can be written as:

$$\frac{1}{\sin\theta}\frac{d}{d\theta}\left(\sin\theta\frac{dF}{d\theta}\right)+\left(k^2R^2-\frac{m^2}{\sin^2\theta}\right)F=0. \tag{20.23}$$

Before we continue let us compare this equation with (20.11) for the modes of the drum which we can rewrite as

$$\frac{1}{r}\frac{d}{dr}\left(r\frac{dF}{dr}\right)+\left(k^2-\frac{m^2}{r^2}\right)F=0. \tag{20.11}$$

Note that these equations are identical when we compare r in (20.11) with $\sin\theta$ in (20.23). There is a good reason for this. Suppose that we have a source in the middle of the drum. Then the variable r measures the distance from a point on the drum to the source. This can be compared with the case of waves on a spherical surface that are excited by a source at the north pole. In that case, $\sin\theta$ is a measure of the distance from a point to the source point. The only difference is that $\sin\theta$ enters the equation rather than the true angular distance θ. This is a consequence of the fact that the surface is curved, and this curvature leaves an imprint on the differential equation that the modes satisfy.

Problem b The differential equation (20.11) was reduced in Section 20.2 to the Bessel equation by changing to a new variable $x = kr$. Define a new variable

$$x \equiv \cos\theta \tag{20.24}$$

and show that the differential equation (20.23) is given by

$$\frac{d}{dx}\left[\left(1-x^2\right)\frac{dF}{dx}\right] + \left(k^2R^2 - \frac{m^2}{1-x^2}\right)F = 0. \tag{20.25}$$

The solution of this differential equation is given by the associated Legendre functions $P_l^m(x)$. These functions are described in great detail in Section 9.8 of Butkov [24]. In fact, just like the Bessel equation, the differential equation (20.25) has a solution that is regular as well as a solution $Q_l^m(x)$ that is singular at the point $x = 1$ where $\theta = 0$. However, since the modes are finite everywhere, they are given only by the regular solution $P_l^m(x)$.

The wavenumber k is related to frequency by the relation $k = \omega/c$. At this point it is not clear what k is, hence the eigenfrequencies of the spherical surface are not yet known. It is shown in Section 9.8 of Butkov [24] that:

- The associated Legendre functions are only finite when the wavenumber satisfies

$$k^2R^2 = l(l+1), \tag{20.26}$$

  where l is a positive integer. Using this in (20.25) implies that the associated Legendre functions satisfy the following differential equation:

$$\frac{1}{\sin\theta}\frac{d}{d\theta}\left[\sin\theta\frac{dP_l^m(\cos\theta)}{d\theta}\right] + \left[l(l+1) - \frac{m^2}{\sin^2\theta}\right]P_l^m(\cos\theta) = 0. \tag{20.27}$$

  Seen as a function of x ($= \cos\theta$) this is equivalent to the following differential equation

$$\frac{d}{dx}\left[\left(1-x^2\right)\frac{dP_l^m(x)}{dx}\right] + \left[l(l+1) - \frac{m^2}{1-x^2}\right]P_l^m(x) = 0.$$

- The integer l must be larger than or equal to the absolute value of the angular order m.

Problem c Show that the last condition can also be written as:

$$-l \le m \le l. \tag{20.28}$$

Problem d Derive that the eigenfrequencies of the modes are given by

$$\omega_l = \sqrt{l(l+1)}\frac{c}{R}. \tag{20.29}$$

It is interesting to compare this result with the eigenfrequencies (20.7) of the string. The eigenfrequencies of the string all have the same spacing in frequency, but the eigenfrequencies of the spherical surface are not spaced at the same interval. In musical jargon one would say that the overtones of a string are harmonious: this means that the eigenfrequencies of the overtones are multiples of the eigenfrequency of the ground tone. In contrast, the overtones of a spherical surface are not harmonious.

Problem e Show that for large values of l the eigenfrequencies of the spherical surface have an almost equal spacing.

Problem f The eigenfrequency ω_l depends on the order l, but not on the degree m. For each value of l, the angular degree m can according to (20.28) take the values $-l, -l+1, \ldots, l-1, l$. Show that this implies that for every value of l, there are $(2l+1)$ modes with the same eigenfrequency.

When different modes have the same eigenfrequency one speaks of *degenerate modes*.

The Legendre functions $P_l^0(x)$ for several degrees l are shown in Figure 20.3 as a function of the variable x. The value $x = 1$ corresponds to $\cos\theta = 1$ or $\theta = 0$: this is the north pole of the spherical coordinate system. The value $x = -1$ corresponds to the south pole of the spherical coordinate system. The number of oscillations of these functions increases with the degree l. All Legendre functions $P_l^0(x)$ have the same value at the north pole: $P_l^0(x = 1) = 1$.

The results we obtained imply that the modes on a spherical surface are given by $P_l^m(\cos\theta)e^{im\varphi}$. We used here that the variable x is related to the angle θ through (20.24). The modes of the spherical surface are called *spherical harmonics*. These

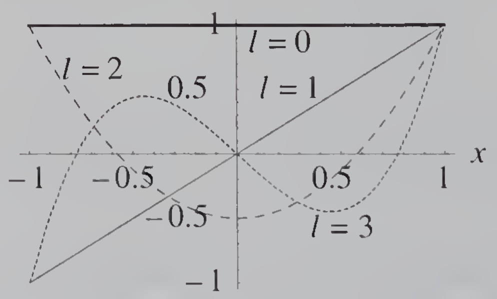

Fig. 20.3 The Legendre polynomials $P_l^0(x)$ for several degrees l.

eigenfunctions for $m \geq 0$ are given by:

$$Y_{lm}(\theta,\varphi) = (-1)^m \sqrt{\frac{2l+1}{4\pi}\frac{(l-m)!}{(l+m)!}} P_l^m(\cos\theta)e^{im\varphi} \qquad m \geq 0. \tag{20.30}$$

For $m < 0$ the spherical harmonics are defined by the relation

$$Y_{lm}(\theta,\varphi) = (-1)^m \, Y_{l,-m}(\theta,\varphi). \tag{20.31}$$

You may wonder where the square-root in front of the associated Legendre function comes from. One can show that with this numerical factor the spherical harmonics are normalized when integrated over the sphere:

$$\iint |Y_{lm}|^2 \, d\Omega = 1, \tag{20.32}$$

where $\iint \cdots d\Omega$ denotes an integration over the unit sphere. You should be aware of the fact that different authors use different definitions of the spherical harmonics. For example, one could also define the spherical harmonics as $\tilde{Y}_{lm}(\theta,\varphi) = P_l^m(\cos\theta)e^{im\varphi}$ because the functions also account for the normal modes of a spherical surface.

Problem g Show that the modes defined in this way satisfy $\iint \left|\tilde{Y}_{lm}\right|^2 d\Omega = 4\pi/(2l+1) \times (l+m)!/(l-m)!$.

This means that the modes defined in this way are not normalized when integrated over the sphere. There is no reason why one cannot work with this convention, as long as one accounts for the fact that in this definition the modes are not normalized. Throughout this book we will use the definition (20.30) for the spherical harmonics. In doing so we follow the normalization that is used by Edmonds [37].

The real parts of the lowest order spherical harmonics are shown in Figure 20.4, in which these functions are projected on the Earth's surface. Only those spherical harmonics with $m \geq 0$ are shown, the spherical harmonics for $m < 0$ follow from (20.31). Just as with the Bessel functions, the associated Legendre functions satisfy recursion relations and a large number of other properties which are described in detail in Section 9.8 of Butkov [24]. The most important properties of the spherical harmonics $Y_{lm}(\theta,\varphi)$ are:

- These functions display m oscillations when the angle φ increases by 2π. In other words, there are m oscillations along one circle of constant latitude.
- The associated Legendre functions $P_l^m(\cos\theta)$ behave like Bessel functions in the sense that they behave like standing waves with an amplitude that decays from the pole. We return to this issue in Section 20.6.

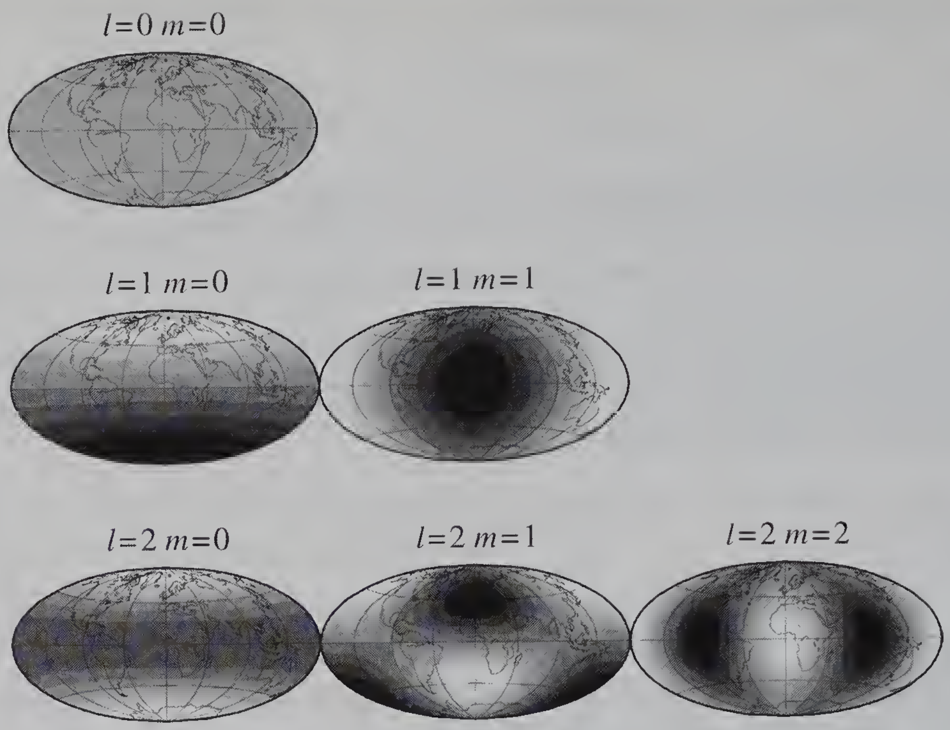

Fig. 20.4 The real parts of the lowest order spherical harmonics. The value of the order l and degree m is shown above each panel. Negative values are shown in white, positive values in black. Each panel shows the surface of the complete sphere using a Hammer projection. The world map is shown to provide spatial orientation. The thin solid lines correspond to lines of constant latitude and longitude on the sphere. Figure courtesy of Jeannot Trampert.

- The number of oscillations between the north pole of the sphere and the south pole of the sphere increases with the difference $(l - m)$.
- The spherical harmonics are orthogonal for a suitably chosen inner product; this orthogonality relation is derived in Section 20.4.
- The spherical harmonics are the eigenfunctions of the Laplacian on the sphere.

Problem h Give a proof of this last property by showing that

$$\nabla_1^2 Y_{lm}(\theta, \varphi) = -l\,(l+1)\,Y_{lm}(\theta, \varphi), \tag{20.33}$$

where the Laplacian on the unit sphere is given by

$$\nabla_1^2 = \frac{1}{\sin\theta}\frac{\partial}{\partial\theta}\left(\sin\theta\frac{\partial}{\partial\theta}\right) + \frac{1}{\sin^2\theta}\frac{\partial^2}{\partial\varphi^2}. \tag{20.34}$$

This property is extremely useful in many applications, because the action of the Laplacian on a sphere can be replaced by the much simpler multiplication by the constant $-l\,(l+1)$ when spherical harmonics are concerned.

20.4 Normal modes of orthogonality relations

The normal modes of a physical system often satisfy orthogonality relations when a suitably chosen inner product for the eigenfunctions is used. In this section this is illustrated by studying once again the normal modes of the Helmholtz equation (20.1) for different geometries. In this section we derive first the general orthogonality relation for these normal modes. This is then applied to the normal modes of the previous sections to derive the orthogonality relations for Bessel functions and associated Legendre functions.

Let us consider two normal modes of the Helmholtz equation (20.1), and let these modes be called u_p and u_q. At this point we leave it open whether the modes are defined on a line, on a surface of arbitrary shape or in a volume. The integration over the region of space in which the modes are defined is denoted as $\int \cdots d^N x$, where N is the dimension of this space. The wavenumbers of these modes, which are effectively the corresponding eigenvalues of the Helmholtz equation, are defined by k_p and k_q, respectively. In other words, the modes satisfy the equations:

$$\nabla^2 u_p + k_p^2 u_p = 0, \tag{20.35}$$

$$\nabla^2 u_q + k_q^2 u_q = 0. \tag{20.36}$$

The subscript p may stand for a single mode index such as in the index n for the wavenumber k_n for the modes of a string, or it may stand for a number of indices such as the indices nm that label the eigenfunctions (20.19) of a circular drum.

Problem a Multiply (20.35) by u_q^*, take the complex conjugate of (20.36) and multiply the result by u_p. Subtract the resulting equations and integrate this over the region of space for which the modes are defined to show that

$$\int \left(u_q^* \nabla^2 u_p - u_p \nabla^2 u_q^* \right) d^N x + \left(k_p^2 - k_q^{*2} \right) \int u_q^* u_p d^N x = 0. \tag{20.37}$$

Problem b Use the theorem of Gauss to derive that

$$\int u_q^* \nabla^2 u_p d^N x = \oint u_q^* \nabla u_p \cdot d\mathbf{S} - \int \left(\nabla u_q^* \cdot \nabla u_p \right) d^N x, \tag{20.38}$$

where the integral $\oint \cdots d\mathbf{S}$ is over the surface that bounds the body. If you have trouble deriving this, you can consult (8.10) where a similar result was used for the derivation of the representation theorem for acoustic waves.

Problem c Use the last result to show that expression (20.37) can be written as

$$\oint \left(u_q^* \nabla u_p - u_p \nabla u_q^*\right) \cdot d\mathbf{S} + \left(k_p^2 - k_q^{*2}\right) \int u_q^* u_p d^N x = 0. \tag{20.39}$$

Problem d The first term is an integral over the boundary of the body. The second term contains a volume integral and this term leads to the orthogonality relation of the modes. Let us assume that on this boundary the modes satisfy one of the three boundary conditions: (i) $u = 0$, (ii) $\hat{\mathbf{n}} \cdot \nabla u = 0$ (where $\hat{\mathbf{n}}$ is the unit vector perpendicular to the surface) or (iii) $\hat{\mathbf{n}} \cdot \nabla u = \alpha u$ (where α is a constant). Show that for all of these boundary conditions the surface integral in (20.39) vanishes.

The last result implies that when the modes satisfy one of these boundary conditions

$$\left(k_p^2 - k_q^{*2}\right) \int u_q^* u_p d^N x = 0. \tag{20.40}$$

Let us first consider the case in which the modes are equal, that is in which $p = q$. In that case the integral reduces to $\int \left|u_p\right|^2 d^N x$ which is guaranteed to be positive. Equation (20.40) then implies that $k_p^2 = k_p^{*2}$, so that the wavenumbers k_p must be real: $k_p = k_p^*$. For this reason the complex conjugate of the wavenumbers can be dropped and (20.40) can be written as:

$$\left(k_p^2 - k_q^2\right) \int u_q^* u_p d^N x = 0. \tag{20.41}$$

Now consider the case of two different modes for which the wavenumbers k_p and k_q are different. In that case the term $\left(k_p^2 - k_q^2\right)$ is nonzero, hence in order to satisfy (20.41) the modes must satisfy

$$\int u_q^* u_p d^N x = 0 \qquad \text{for} \quad k_p \neq k_q. \tag{20.42}$$

This finally gives the *orthogonality relation* of the modes in the sense that it states that the modes are orthogonal for the following inner product: $\langle f \cdot g \rangle \equiv \int f^* g \, d^N x$. Note that the inner product for which the modes are orthogonal follows from the Helmholtz equation (20.1) which defines the modes.

Let us now consider this orthogonality relation for the modes of the string, the drum, and the spherical surface of the previous sections. For the string, the orthogonality relation was derived in problem d of Section 20.1 and you can see that equation (20.6) is identical to the general orthogonality relation (20.42). For the circular drum the modes are given by (20.19).

Problem e Use (20.19) for the modes of the circular drum to show that the orthogonality relation (20.42) for this case can be written as:

$$\int_0^R \int_0^{2\pi} J_{m_1}(k_{n_1}^{(m_1)} r) J_{m_2}(k_{n_2}^{(m_2)} r) e^{i(m_1 - m_2)\varphi} d\varphi \, r dr$$
$$= 0 \quad \text{for} \quad k_{n_1}^{(m_1)} \neq k_{n_2}^{(m_2)}. \tag{20.43}$$

Explain where the factor r comes from in the integration.

Problem f This integral can be separated into an integral over φ and an integral over r. The φ-integral is given by $\int_0^{2\pi} e^{i(m_1-m_2)\varphi} d\varphi$. Show that this integral vanishes when $m_1 \neq m_2$:

$$\int_0^{2\pi} e^{i(m_1-m_2)\varphi} d\varphi = 0 \quad \text{for} \quad m_1 \neq m_2. \tag{20.44}$$

Note that you obtained this relation earlier in (17.9) in the derivation of the residue theorem.

Expression (20.44) implies that the modes $u_{n_1 m_1}(r, \varphi)$ and $u_{n_2 m_2}(r, \varphi)$ are orthogonal when $m_1 \neq m_2$ because the φ-integral in (20.43) vanishes when $m_1 \neq m_2$. Let us now consider why the different modes of the drum are orthogonal when m_1 and m_2 are equal to the same integer m. In that case (20.43) implies that

$$\int_0^R J_m(k_{n_1}^{(m)} r) J_m(k_{n_2}^{(m)} r) \, r \, dr = 0 \quad \text{for} \quad n_1 \neq n_2. \tag{20.45}$$

Note that we have used here that $k_{n_1}^{(m)} \neq k_{n_2}^{(m)}$ when $n_1 \neq n_2$. This integral defines an orthogonality relation for Bessel functions. Note that both Bessel functions in this relation are of the same degree m but that the wavenumbers in the argument of the Bessel functions differ. Note the resemblance between this expression and the orthogonality relation (20.6) of the modes of the string which can be written as

$$\int_0^{2R} \sin(k_n x) \, \sin(k_m x) \, dx = 0 \quad \text{for} \quad n \neq m. \tag{20.46}$$

The presence of the term r in the integral (20.45) comes from the fact that the modes of the drum are orthogonal for the integration over the total area of the drum. In cylindrical coordinates this leads to a factor r in the integration.

Problem g Take your favorite book on mathematical physics and find an alternative derivation of the orthogonality relation (20.45) of the Bessel functions of the same degree m.

Note finally that the modes $u_{n_1m_1}(r,\varphi)$ and $u_{n_2m_2}(r,\varphi)$ are orthogonal when $m_1 \neq m_2$ because the φ-integral satisfies (20.44), whereas the modes are orthogonal when $n_1 \neq n_2$ but with the same order m because the r-integral (20.45) vanishes in that case. This implies that the eigenfunctions of the drum defined in (20.19) satisfy the following orthogonality relation:

$$\int_0^R \int_0^{2\pi} u^*_{n_1m_1}(r,\varphi)u_{n_2m_2}(r,\varphi)\, d\varphi\; r dr = C\delta_{n_1n_2}\delta_{m_1m_2}, \tag{20.47}$$

where δ_{ij} is the Kronecker delta and C is a constant that depends on n_1 and m_1.

A similar analysis can be applied to the spherical harmonics $Y_{lm}(\theta,\varphi)$ which are the eigenfunctions of the Helmholtz equation on a spherical surface. You may wonder in that case what the boundary conditions of these eigenfunctions are because in the step from (20.39) to (20.40) the boundary conditions of the modes have been used. A closed surface, however, has no boundary. This means that the surface integral in (20.39) vanishes. This in turn means that the orthogonality relation (20.42) holds despite the fact that the spherical harmonics do not satisfy one of the boundary conditions that was used in problem d. Let us now consider the inner product of two spherical harmonics on the sphere: $\iint Y^*_{l_1m_1}(\theta,\varphi)Y_{l_2m_2}(\theta,\varphi)d\Omega$.

Problem h Show that the φ-integral in the integration over the sphere is of the form $\int_0^{2\pi} e^{i(m_2-m_1)}d\varphi$ and that this integral is equal to $2\pi\delta_{m_1m_2}$.

This implies that the spherical harmonics are orthogonal when $m_1 \neq m_2$ because of the φ-integration. We will now continue with the case in which $m_1 = m_2$, and denote this common value with the single index m.

Problem i Use the general orthogonality relation (20.42) to derive that the associated Legendre functions satisfy the following orthogonality relation:

$$\int_0^{\pi} P^m_{l_1}(\cos\theta)P^m_{l_2}(\cos\theta)\sin\theta d\theta = 0 \qquad \text{when } l_1 \neq l_2. \tag{20.48}$$

Note the common value of the degree m in the two associated Legendre functions. Also show explicitly that the condition $k_{l_1} \neq k_{l_2}$ is equivalent to the condition $l_1 \neq l_2$.

Problem j Use a substitution of variables to show that this orthogonality relation can also be written as

$$\int_{-1}^{1} P^m_{l_1}(x)P^m_{l_2}(x)dx = 0 \qquad \text{when } l_1 \neq l_2. \tag{20.49}$$

Problem k Find an alternative derivation of this orthogonality relation in the literature.

The result you obtained in problem h implies that the spherical harmonics are orthogonal when $m_1 \neq m_2$ because of the φ-integration, whereas problem i implies that the spherical harmonics are orthogonal when $l_1 \neq l_2$ because of the θ-integration. This means that the spherical harmonics satisfy the following orthogonality relation:

$$\iint Y^*_{l_1 m_1}(\theta, \varphi) Y_{l_2 m_2}(\theta, \varphi)\, d\Omega = \delta_{l_1 l_2} \delta_{m_1 m_2}. \tag{20.50}$$

The numerical constant multiplying the delta functions is equal to 1. This is a consequence of the square-root term in (20.30) that pre-multiplies the associated Legendre functions. Be aware of the fact that when a different convention is used for the normalization of the spherical harmonics a normalization factor appears on the right-hand side of the orthogonality relation (20.50) of the spherical harmonics.

20.5 Bessel functions behave as decaying cosines

As we have seen in Section 20.2 the modes of the circular drum are given by $J_m(kr)e^{im\varphi}$, where the Bessel function satisfies the differential equation (20.16) and where k is a wavenumber chosen in such a way that the displacement at the edge of the drum vanishes. We show in this section that the waves that propagate through the drum have an approximately constant wavelength, but that their amplitude decays with the distance to the center of the drum. The starting point of the analysis is the Bessel equation

$$\frac{d^2 J_m}{dx^2} + \frac{1}{x}\frac{dJ_m}{dx} + \left(1 - \frac{m^2}{x^2}\right) J_m = 0. \tag{20.16}$$

If the terms $(1/x)dJ_m/dx$ and m^2/x^2 were absent in (20.16) the Bessel equation would reduce to the differential equation $d^2F/dx^2 + F = 0$ whose solutions are given by a superposition of $\cos x$ and $\sin x$. We can therefore expect the Bessel functions to display an oscillatory behavior when x is large.

It follows directly from (20.16) that the term m^2/x^2 is relatively small for large values of x, specifically when $x \gg m$. However, it is not obvious under which conditions the term $(1/x)dJ_m/dx$ is relatively small. Fortunately this term can be transformed away.

Problem a Write $J_m(x) = x^\alpha g_m(x)$, insert this in the Bessel equation (20.16), and show that the term with the first derivative vanishes when $\alpha = -1/2$ and that the resulting differential equation for $g_m(x)$ is given by

$$\frac{d^2 g_m}{dx^2} + \left(1 - \frac{m^2 - 1/4}{x^2}\right) g_m = 0. \tag{20.51}$$

Up to this point we have made no approximations. Although we have transformed the first derivative term out of the Bessel equation, we still cannot solve (20.51). However, when $x \gg m$ the term proportional to $1/x^2$ in this expression is relatively small. This means that for large values of x the function $g_m(x)$ satisfies the approximate differential equation $d^2 g_m/dx^2 + g_m \approx 0$.

Problem b Show that the solution of this equation is given by $g_m(x) \approx A\cos(x + \varphi)$, where A and φ are constants. Also show that this implies that the Bessel function is approximately given by:

$$J_m(x) \approx A\frac{\cos(x + \varphi)}{\sqrt{x}}. \tag{20.52}$$

This approximation is obtained from a local analysis of the Bessel equation. Since all values of the constants A and φ lead to a solution that approximately satisfies the differential equation (20.51), it is not possible to retrieve the precise values of these constants from the analysis in this section. An analysis based on the asymptotic evaluation of the integral representation of the Bessel function as presented in Section 24.5 or reference [14] shows that:

$$J_m(x) = \sqrt{\frac{2}{\pi x}} \cos\left[x - (2m + 1)\frac{\pi}{4}\right] + O(x^{-3/2}). \tag{20.53}$$

Problem c As a check on the accuracy of this asymptotic expression let us compare the zeroes of this approximation with the zeroes of the Bessel functions as given in Table 20.1. In problem h of Section 20.2 you found that the separation of the zero-crossings tends to π for large values of x. Explain this using the approximate expression (20.53). How large must x be for the different values of the degree m so that the difference in the spacing of the zero-crossings with π is less than 0.01?

The asymptotic expression (20.53) can be compared in Figure 20.5 with the Bessel functions $J_m(x)$ for the degrees $m = 0$ and $m = 2$, respectively. For large values of x, the approximation (20.53) is very good. Note that for $m = 0$ the approximation is better for smaller values of x than it is for $m = 2$. This is related to the fact that the approximation (20.53) is valid under the condition

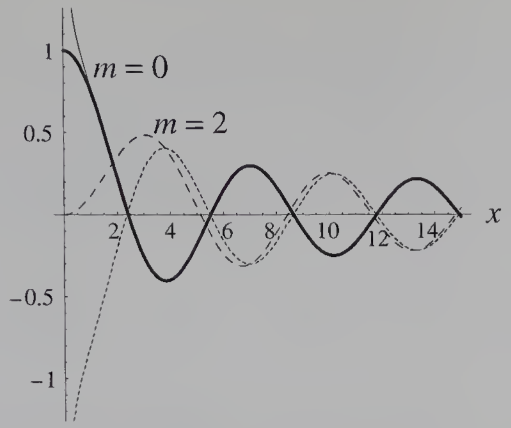

Fig. 20.5 The Bessel functions $J_0(x)$ (thick solid line) and $J_2(x)$ (dashed line) and their approximations (thin solid line and dotted line respectively). Note that for $m = 0$ the approximation is better for smaller values of x than it is for $m = 2$.

$\left(m^2 - 1/4\right)/x^2 \ll 1$. This requirement is satisfied when $x \gg m$, hence the approximation (20.53) is for a given value of x better for the smaller degrees m than for larger degrees m.

Physically, (20.53) states that Bessel functions behave like standing waves with a constant wavelength and which decay with distance as $1/\sqrt{kr}$. (Here it is used that the modes are given by the Bessel functions with argument $x = kr$.) How can we explain this decay of the amplitude with distance? First let us note that (20.53) expresses the Bessel function in a cosine, hence this is a representation of the Bessel function as a standing wave. However, using the relation $\cos x = \left(e^{ix} + e^{-ix}\right)/2$ the Bessel function can be written as two traveling waves that depend on the distance r as $e^{\pm ikr}/\sqrt{kr}$. These waves interfere to give the standing wave pattern of the Bessel function. Now let us consider a propagating wave $A(r)e^{ikr}$ in two dimensions; in this expression $A(r)$ is an amplitude that is at this point unknown. The energy of the wave varies with the square of the wave field, and thus depends on $|A(r)|^2$. The energy current therefore also varies as $|A(r)|^2$. Consider an outgoing wave as shown in Figure 20.6. The total energy flux through a ring of radius r is given by the energy current times the circumference of the ring, which means that the flux is equal to $2\pi r\,|A(r)|^2$. Since energy is conserved, this total energy flux is the same for all values of r, which means that $2\pi r\,|A(r)|^2 = \text{constant}$.

Problem d Show that this implies that $A(r) \sim 1/\sqrt{r}$.

This is the same dependence on distance as the $1/\sqrt{x}$ decay of the approximation (20.53) of the Bessel function. This means that the decay of the Bessel function with distance is dictated by the requirement of energy conservation.

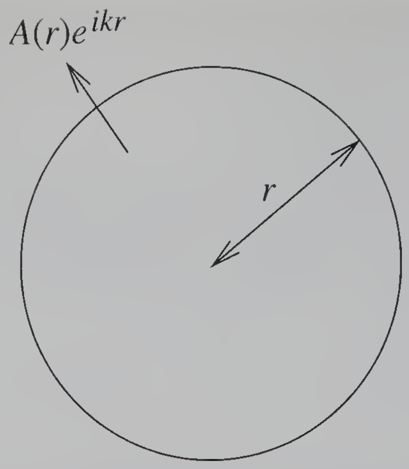

Fig. 20.6 An expanding wavefront with radius r.

20.6 Legendre functions behave as decaying cosines

The technique used in the previous section for the approximation of the Bessel function can also be applied to spherical harmonics. We show in this section that the spherical harmonics behave asymptotically as standing waves on a sphere with an amplitude decay that is determined by the condition that energy is conserved. The spherical harmonics are proportional to the associated Legendre functions with argument $\cos\theta$. The starting point of our analysis therefore is the differential equation for $P_l^m(\cos\theta)$ that was derived in Section 20.3:

$$\frac{1}{\sin\theta}\frac{d}{d\theta}\left[\sin\theta\frac{dP_l^m(\cos\theta)}{d\theta}\right]+\left[l\,(l+1)-\frac{m^2}{\sin^2\theta}\right]P_l^m(\cos\theta)=0. \quad (20.27)$$

Let us assume that we have a source at the north pole, where $\theta = 0$. Far away from the source, the term $m^2/\sin^2\theta$ in the last term on the left-hand side is much smaller than the constant $l\,(l+1)$.

Problem a Show that the words "far away from the source" stand for the requirement

$$\sin\theta \gg \frac{m}{\sqrt{l\,(l+1)}}, \quad (20.54)$$

and show that this implies that the approximation that we derive breaks down near the north pole as well as near the south pole of the employed system of spherical coordinates. In addition, the asymptotic expressions that we derive are most accurate for large values of the angular order l and small values of the degree m.

Problem b Just as in the previous section we can transform the first derivative in the differential equation (20.27) away: here this can be achieved by writing $P_l^m(\cos\theta) = (\sin\theta)^\alpha\, g_l^m(\theta)$. Insert this substitution into the differential equation (20.27) and show that the first derivative $dg_l^m/d\theta$ disappears when $\alpha = -1/2$ and that the resulting differential equation for $g_l^m(\theta)$ is given by:

$$\frac{d^2 g_l^m}{d\theta^2} + \left[\left(l + \frac{1}{2}\right)^2 - \frac{m^2 - 1/4}{\sin^2\theta}\right] g_l^m(\theta) = 0. \tag{20.55}$$

It is interesting to note the resemblance of this equation to the corresponding expression (20.51) in the analysis of the Bessel function.

Problem c If the term $\left(m^2 - 1/4\right)/\sin^2\theta$ were absent, this equation would be simple to solve. Show that this term is small compared to the constant $(l + \frac{1}{2})^2$ when the requirement (20.54) is satisfied.

Problem d Show that under this condition the associated Legendre functions satisfy the following approximation:

$$P_l^m(\cos\theta) \approx A \frac{\cos\left[\left(l + \frac{1}{2}\right)\theta + \gamma\right]}{\sqrt{\sin\theta}}, \tag{20.56}$$

where A and γ are constants.

Just as in the previous section the constants A and γ cannot be obtained from this analysis because (20.56) satisfies the approximate differential equation for *any* values of these constants. As shown in (2.5.58) of reference [37] the asymptotic relation of the associated Legendre functions is given by:

$$P_l^m(\cos\theta) \approx (-l)^m \sqrt{\frac{2}{\pi l \sin\theta}} \cos\left[\left(l + \frac{1}{2}\right)\theta - (2m+1)\frac{\pi}{4}\right] + O(l^{-3/2}). \tag{20.57}$$

Just like the Bessel functions the spherical harmonics behave like a standing wave given by a cosine that is multiplied by a factor $1/\sqrt{\sin\theta}$ which modulates the amplitude.

Problem e Use the same reasoning as in problem d of Section 20.5 to explain that this amplitude decrease follows from the requirement of energy conservation. In doing so you may find Figure 20.7 helpful.

Problem f Deduce from (20.57) that the wavelength of the associated Legendre functions measured in radians is given by $2\pi/(l + \frac{1}{2})$.

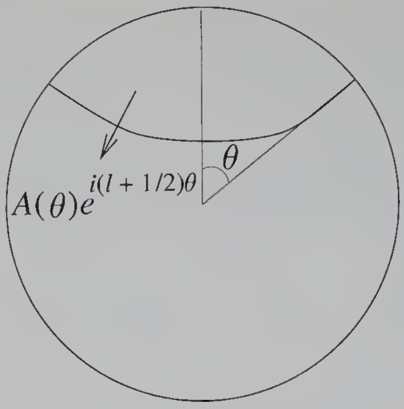

Fig. 20.7 An expanding wavefront on a spherical surface at a distance θ from the source.

This last result can be used to find the number of oscillations in the spherical harmonics when one moves around the globe once. For simplicity we consider here the case of a spherical harmonic $Y_l^0(\theta, \varphi)$ for degree $m = 0$. When one goes from the north pole to the south pole, the angle θ increases from 0 to π. The number of oscillations which fit in this interval is given by $\pi/\textit{wavelength}$, and according to problem f this number is equal to $\pi/[2\pi/(l + \frac{1}{2})] = (l + \frac{1}{2})/2$. This is the number of wavelengths that fit on half the globe. When one returns from the south pole to the north pole one encounters another $(l + \frac{1}{2})/2$ oscillations. This means that the total number of waves that fit around the globe is given by $(l + \frac{1}{2})$. It may surprise you that the number of oscillations that one encounters making one loop around the globe is not an integer. One would expect that the requirement of constructive interference would dictate that an integer number of wavelengths should "fit" in this interval. The reason why the total number of oscillations is $(l + \frac{1}{2})$ rather than the integer l is that near the north and south poles the asymptotic approximation (20.57) breaks down; this follows from the requirement (20.54).

The fact that $(l + \frac{1}{2})$ rather than l oscillations fit on the sphere has a profound effect in quantum mechanics. In the first attempts to explain the line spectra of light emitted by atoms, Bohr postulated that an integer number of waves has to fit on a sphere: this can be expressed as $\oint k ds = 2\pi n$, where k is the local wave-number. This condition could not explain the observed spectra of light emitted by atoms. However, the arguments in this section imply that the number of wavelengths that fit on a sphere should be given by the requirement

$$\oint k ds = 2\pi \left(n + \frac{1}{2} \right) . \tag{20.58}$$

This is the *Bohr–Sommerfeld quantization rule*, and was the earliest result in quantum mechanics which provided an explanation of the line spectra of light emitted by atoms. More details on this issue and the cause of the factor $\frac{1}{2}$ in the quantization rule can be found in references [104] and [20]. The effect of this factor on the Earth's normal modes is discussed in a pictorial way by Dahlen and Henson [32].

The asymptotic expression (20.57) can give a useful insight into the relation between modes and traveling waves on a sphere. Let us first return to the modes on the string, which according to (20.5) are given by $\sin(k_n x)$. For simplicity, we will leave out normalization constants in the arguments. The wave motion associated with this mode is given by the real part of $\sin(k_n x)e^{-i\omega_n t}$, with $\omega_n = k_n/c$. These modes therefore denote a standing wave. However, using the decomposition $\sin(k_n x) = (e^{ik_n x} - e^{-ik_n x})/2i$, the mode can in the time domain also be seen as a superposition of two waves $e^{i(k_n x - \omega_n t)}$ and $e^{-i(k_n x + \omega_n t)}$. These are two traveling waves which move in opposite directions.

Problem g Suppose we excite a string at the left-hand side at $x = 0$. We know we can account for the motion of the string as a superposition of standing waves $\sin(k_n x)$. However, we can consider these modes to exist also as a superposition of waves $e^{\pm ik_n x}$ which move in opposite directions. The wave $e^{ik_n x}$ moves away from the source at $x = 0$. However the wave $e^{-ik_n x}$ moves *towards* the source at $x = 0$. Give a physical explanation of why in the string traveling waves also move *towards* the source.

On a sphere the situation is analogous. The modes can be written according to (20.57) as standing waves

$$\cos\left[\left(l + \frac{1}{2}\right)\theta - (2m+1)\frac{\pi}{4}\right] / \sqrt{\sin\theta}$$

on the sphere. Using the relation $\cos x = \left(e^{ix} + e^{-ix}\right)/2$ the modes can also be seen as a superposition of traveling waves $e^{i(l+\frac{1}{2})\theta}/\sqrt{\sin\theta}$ and $e^{-i(l+\frac{1}{2})\theta}/\sqrt{\sin\theta}$ on the sphere.

Problem h Explain why the first wave travels away from the north pole while the second wave travels towards the north pole.

Problem i Suppose that the waves are excited by a source at the north pole. According to the last problem the motion of the sphere can alternatively be seen as a superposition of standing waves or of traveling waves. The traveling wave $e^{i(l+\frac{1}{2})\theta}/\sqrt{\sin\theta}$ moves away from the source. Explain physically why there is also a traveling wave $e^{-i(l+\frac{1}{2})\theta}/\sqrt{\sin\theta}$ moving *towards* the source.

These results imply that the oscillations of the Earth can be seen as either a superposition of normal modes, or a superposition of waves that travel along the Earth's surface in opposite directions. The waves that travel along the Earth's surface are called *surface waves*. The relation between normal modes and surface waves is treated in more detail by Dahlen [31] and by Snieder and Nolet [101].

20.7 Normal modes and the Green's function

In Section 13.6 we analyzed the normal modes of a system of three coupled masses. This system had three normal modes, and each mode could be characterized by a vector $\hat{\mathbf{v}}^{(n)}$ with the displacement of the three masses and by an eigenfrequency ω_n. The response of the system to a force $\mathbf{F}$ acting on the three masses with time dependence $e^{-i\omega t}$ was derived to be:

$$\mathbf{x} = \frac{1}{m}\sum_{n=1}^{3}\frac{\hat{\mathbf{v}}^{(n)}(\hat{\mathbf{v}}^{(n)}\cdot\mathbf{F})}{\left(\omega_n^2-\omega^2\right)}. \tag{13.67}$$

This means that the Green's function of this system is given by the following dyad:

$$\mathbf{G} = \frac{1}{2\pi m}\sum_{n=1}^{3}\frac{\hat{\mathbf{v}}^{(n)}\hat{\mathbf{v}}^{(n)T}}{\left(\omega_n^2-\omega^2\right)}. \tag{20.59}$$

The factor $1/2\pi$ is due to the fact that a delta-function force $f(t)=\delta(t)$ in the time domain corresponds with the Fourier transform (15.43) to $F(\omega)=1/2\pi$ in the frequency domain. In this section we derive the Green's function for a general oscillating system that can be continuous. An important example is the Earth, which is a body that has well-defined normal modes and where the displacement is a continuous function of the space coordinates.

We consider a system that satisfies the following equation of motion:

$$\rho\ddot{u} + Hu = F. \tag{20.60}$$

The field u can be either a scalar field or a vector field. The operator H at this point is general, the only requirement that we impose is that this operator is Hermitian, which means that we require that

$$(f\cdot Hg) = (Hf\cdot g)\,, \tag{20.61}$$

where the inner product is defined as $(f\cdot h)\equiv\int f^{*}g\,dV$. In the frequency domain, the equation of motion is given by

$$-\rho\omega^2 u + Hu = F(\omega). \tag{20.62}$$

Let the normal modes of the system be denoted by $u^{(n)}$; the normal modes describe the oscillations of the system in the absence of any external force. The normal modes therefore satisfy the following expression

$$H u^{(n)} = \rho \omega_n^2 u^{(n)}, \tag{20.63}$$

where ω_n is the eigenfrequency of this mode.

Problem a Take the inner product of this expression with a mode $u^{(m)}$, and use the fact that H is Hermitian to derive that

$$\left(\omega_n^2 - \omega_m^{*2}\right)\left(u^{(m)} \cdot \rho u^{(n)}\right) = 0. \tag{20.64}$$

Note the resemblance of this expression to (20.40) for the modes of a system that obeys the Helmholtz equation.

Problem b Just like in Section 20.4 one can show that the eigenfrequencies are real by setting $m = n$, and one can derive that different modes are orthogonal with respect to the following inner product:

$$\left(u^{(m)} \cdot \rho u^{(n)}\right) = \delta_{nm} \qquad \text{for} \quad \omega_m \neq \omega_n. \tag{20.65}$$

Use (20.64) to give a proof of this orthogonality relation.

Note the presence of the density term ρ in this inner product.

In the analysis of this section it was crucial for the operator H to be Hermitian. This property of H has two important implications: (1) the eigenvalues of H are real and (2) the eigenfunctions are orthogonal. It is for this reason that it is crucial to establish whether the operator H is Hermitian or not. In general, the operator of a dynamical system is Hermitian when the system is invariant for time-reversal. This means that the equations are invariant when one lets the clock run backward, or mathematically when one replaces t by $-t$. Dissipation in general breaks the symmetry for time-reversal. It is shown in detail by Dahlen and Tromp [33] that attenuation in the Earth makes the eigenfrequencies of the Earth complex and that the normal modes of an attenuating Earth do not satisfy the orthogonality relation (20.65).

Let us now return to the inhomogeneous problem (20.62) where an external force $F(\omega)$ is present. Assuming that the normal modes form a complete set, the response to this force can be written as a sum of normal modes:

$$u = \sum_n c_n u^{(n)}, \tag{20.66}$$

where the c_n are unknown coefficients.

Problem c Find these coefficients by inserting (20.66) into the equation of motion (20.62) and by taking the inner product of the result with a mode $u^{(m)}$ to derive that

$$c_m = \frac{\left(u^{(m)} \cdot F\right)}{\omega_m^2 - \omega^2}. \tag{20.67}$$

This means that the response of the system can be written as:

$$u = \sum_n \frac{u^{(n)}\left(u^{(n)} \cdot F\right)}{\omega_n^2 - \omega^2}. \tag{20.68}$$

Note the resemblance of this expression to (13.67) for a system of three masses. The main difference is that the derivation in this section is also valid for continuous vibrating systems such as the Earth.

It is instructive to rewrite this expression taking the dependence of the space coordinates explicitly into account:

$$u(\mathbf{r}) = \sum_n \frac{u^{(n)}(\mathbf{r}) \int u^{*(n)}(\mathbf{r}')F(\mathbf{r}')\,dV'}{\omega_n^2 - \omega^2}. \tag{20.69}$$

It follows from this expression that the Green's function is given by

$$G(\mathbf{r}, \mathbf{r}', \omega) = \frac{1}{2\pi} \sum_n \frac{u^{(n)}(\mathbf{r})u^{*(n)}(\mathbf{r}')}{\omega_n^2 - \omega^2}. \tag{20.70}$$

When the mode is a vector, one should take the transpose of the mode $u^{*(n)}(\mathbf{r}')$. Note the similarity between this expression for the Green's function of a continuous medium and the Green's function (20.59) for a discrete system. In this sense, the Earth behaves in the same way as a tri-atomic molecule. For both systems, the dyadic representation of the Green's function provides a compact way to account for the response of the system to external forces.

Note that the response is strongest when the frequency ω of the external force is close to one of the eigenfrequencies ω_n of the system. This implies for example for the Earth that modes with a frequency close to the frequency of the external forcing are most strongly excited. If we jump up and down with a frequency of 1 Hz, we excite the Earth's fundamental mode with a period of about 1 hour only very weakly. In addition, a mode is most effectively excited when the inner product of the forcing $F(\mathbf{r}')$ in (20.69) is maximal. This means that a mode is most strongly excited when the spatial distribution of the force equals the displacement $u^{(n)}(\mathbf{r}')$ of the mode.

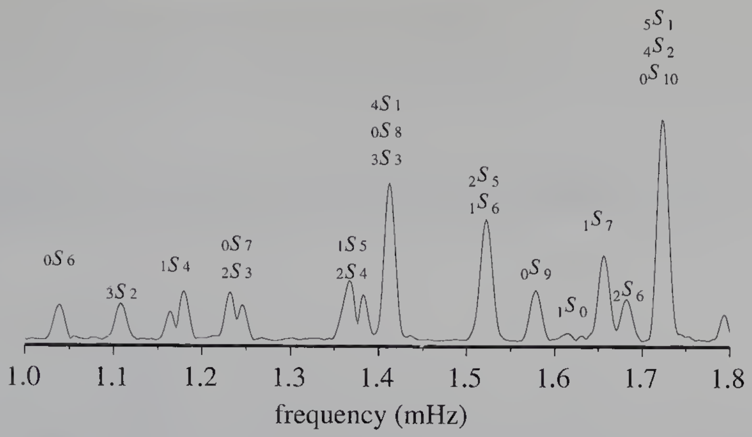

Fig. 20.8 Amplitude spectrum of the vertical component of the ground motion at a seismic station in Tucson, Arizona after the 9 June 1994 Bolivian earthquake. The numbers $_nS_l$ denote the different normal modes of the Earth. Figure courtesy of Arwen Deuss.

Problem d Show that a mode is not excited when the force acts only at one of the nodal lines of that mode.

The normal modes of the Earth leave a clear imprint of the motion of the Earth after a strong earthquake. The ground motion after the 9 June 1994 earthquake in Bolivia was recorded in Tucson, Arizona. Figure 20.8 shows the amplitude spectrum of the vertical component of the ground motion after the earthquake. Note that the frequency is given in units of millihertz because the Earth oscillates slowly. The peaks in the amplitude spectrum correspond to the normal modes of the Earth. These peaks are described by the function $1/\left(\omega_n^2 - \omega^2\right)$ in (20.70). In reality the detailed structure of these resonances is also affected by the attenuation in the Earth and by a number of factors that perturb the Earth's normal modes.

As a next step we consider the Green's function in the time domain. This function follows by applying the Fourier transform (15.42) to the Green's function (20.70).

Problem e Show that this gives:

$$G(\mathbf{r}, \mathbf{r}', t) = \frac{1}{2\pi} \sum_n u^{(n)}(\mathbf{r}) u^{*(n)}(\mathbf{r}') \int_{-\infty}^{\infty} \frac{e^{-i\omega t}}{\omega_n^2 - \omega^2} d\omega. \tag{20.71}$$

The integrand is singular at the frequencies $\omega = \pm\omega_n$ of the normal modes. These singularities are located on the integration path, as shown in the left-hand panel of Figure 20.9. At the singularity at $\omega = \omega_n$ the integrand behaves as

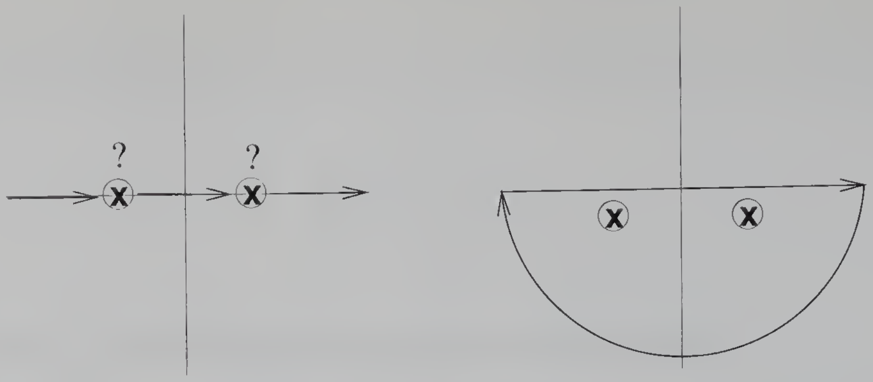

Fig. 20.9 The location of the poles and the integration path in the complex ω-plane. The poles are indicated with a cross. The left-hand panel shows the original situation where the poles are located on the integration path at location $\pm\omega_n$. The right-hand panel shows the location of the poles when a slight anelastic damping is present.

$1/[2\omega_n(\omega - \omega_n)]$. The contribution of these singularities is poorly defined because the integral $\int 1/(\omega - \omega_n)\ d\omega$ is not defined.

This situation is comparable to the treatment in Section 17.4 of the response of a particle in syrup to an external forcing. When this particle was subjected to a damping β, the integrand in the Fourier transform to the time domain had a singularity in the lower half-plane. This gave a causal response; as shown in (17.35) the response was different from zero only at times *later* than the time at which the forcing was applied. This suggests that we can obtain a well-defined causal response of the Green's function (20.71) when we introduce a slight damping. This damping breaks the invariance of the problem for time-reversal, and is responsible for a causal response. At the end of the calculation we can let the damping parameter go to zero. Damping can be introduced by giving the eigenfrequencies of the normal modes a small negative imaginary component: $\pm\omega_n \rightarrow \pm\omega_n - i\eta$, where η is a small positive number.

Problem f The time dependence of the oscillation of a normal mode is given by $e^{-i\omega_n t}$. Show that with this replacement the modes decay with a decay time that is given by

$$\tau = 1/\eta. \qquad (20.72)$$

This last property means that when we ultimately set $\eta = 0$ the decay time becomes infinite: in other words, the modes are not attenuated in that limit.

With the replacement $\pm\omega_n \rightarrow \pm\omega_n - i\eta$ the poles that are associated with the normal modes are located in the lower ω-plane: this situation is shown in

Figure 20.9. Now that the singularities are moved from the integration path, the theory of complex integration can be used to evaluate the resulting integral.

Problem g Use the theory of contour integration as treated in Chapter 17 to derive that the Green's function is given in the time domain by:

$$G(\mathbf{r},\mathbf{r}',t)=\begin{cases}0 & \text{for} \quad t<0\\ \displaystyle\sum_n \frac{u^{(n)}(\mathbf{r})u^{*(n)}(\mathbf{r}')}{\omega_n}\sin\omega_n t & \text{for} \quad t>0\end{cases}. \tag{20.73}$$

Hint: use the same steps as in the derivation of the function (17.35) and let the damping parameter η go to zero at the end of the integration.

This result gives a causal response because the Green's function is only nonzero at times $t>0$, which is later than the time $t=0$ when the delta-function forcing is nonzero. The total response is given as a sum over all the modes. Each mode leads to a time signal $\sin(\omega_n t)$ in the modal sum: this is a periodic oscillation with the frequency ω_n of the mode. The singularities in the integrand of the Green's function (20.71) at the pole positions $\omega=\pm\omega_n$ are thus associated in the time domain with a harmonic oscillation with angular frequency ω_n. Note that the Green's function is continuous at the time $t=0$ of excitation. It is interesting to compare this Green's function with the Green's function (18.21) for the girl on the swing. The only differences are that in (20.73) modes are present in the Green's function and that a summation over the modes is carried out. This difference is due to the fact that the harmonic oscillator has only one mode and that this mode of oscillation does not have a spatial extent.

Problem h Use the Green's function (20.73) to derive that the response of the system to a force $F(\mathbf{r},t)$ is given by:

$$u(\mathbf{r},t)=\sum_n \frac{1}{\omega_n}u^{(n)}(\mathbf{r})\int\int_{-\infty}^{t} u^{*(n)}(\mathbf{r}')\sin\omega_n\left(t-t'\right)F(\mathbf{r}',t')\,dt'dV'. \tag{20.74}$$

Justify the integration limit in the t'-integration.

The results in this section imply that the total Green's function of a system is known once the normal modes are known. The total response can then be obtained by summing the contribution of each normal mode to the total response. This technique is called *normal-mode summation*, and is often used to obtain the low-frequency response of the Earth to an excitation [33, 35]. However, in the seismological literature one usually treats a source signal that is given by a step function at $t = 0$ rather than a delta function because this is a more accurate description of the slip on a fault during an earthquake [3]. This leads to a time dependence $(1 - \cos(\omega_n t))$ rather than the time dependence $\sin(\omega_n t)$ in the response (20.73) to a delta-function excitation.

20.8 Guided waves in a low-velocity channel

In this section we consider a system that strictly speaking does not have normal modes, but that can support solutions that behave like traveling waves in one direction and as modes in another direction. The waves in such a system propagate as *guided waves*. Consider a system in two dimensions (x and z), where the velocity depends only on the z-coordinate. We assume that the wave-field satisfies the Helmholtz equation (20.1) in the frequency domain:

$$\nabla^2 u + \frac{\omega^2}{c^2(z)} u = 0. \tag{20.75}$$

In this section we consider a simple model of a layer of thickness H that extends from $z = 0$ to $z = H$ in which the velocity is given by c_1. This layer is embedded in a medium with a constant velocity c_0. The geometry of the problem is shown in Figure 20.10. Since the system is invariant in the x-direction, the problem can be simplified by a Fourier transform over the x-coordinate:

$$u(x, z) = \int_{-\infty}^{\infty} U(k, z) e^{ikx} dk. \tag{20.76}$$

Problem a Show that $U(k, z)$ satisfies the following ordinary differential equation:

$$\frac{d^2 U}{dz^2} + \left[\frac{\omega^2}{c(z)^2} - k^2 \right] U = 0. \tag{20.77}$$

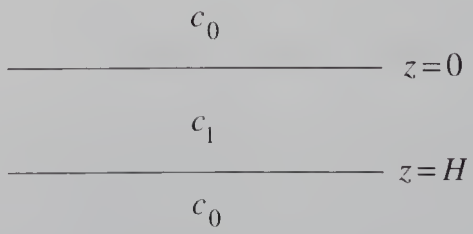

Fig. 20.10 Geometry of the model of a single layer sandwiched between two homogeneous half-spaces.

It is important to note at this point that the frequency ω is a fixed constant, and that according to (20.76) the variable k is an integration variable that assumes all values in the integration (20.76). For this reason one should not at this point use the relation $k = \omega/c(z)$.

Now consider the special case of the model shown in Figure 20.10. We require that the waves outside the layer move away from the layer.

Problem b Show that this implies that the solution for $z < 0$ is given by $Ae^{-ik_0 z}$ and the solution for $z > H$ is given by $Be^{+ik_0 z}$ where A and B are unknown integration constants and where k_0 is given by

$$k_0 = \sqrt{\frac{\omega^2}{c_0^2} - k^2}. \tag{20.78}$$

Problem c Show that within the layer the wave field is given by $C\cos(k_1 z) + D\sin(k_1 z)$ with C and D integration constants and k_1 given by

$$k_1 = \sqrt{\frac{\omega^2}{c_1^2} - k^2}. \tag{20.79}$$

The solution in the three regions of space therefore takes the following form:

$$U(k, z) = \begin{cases} Ae^{-ik_0 z} & \text{for} \quad z < 0 \\ C\cos(k_1 z) + D\sin(k_1 z) & \text{for} \quad 0 < z < H. \\ Be^{+ik_0 z} & \text{for} \quad z > H \end{cases} \tag{20.80}$$

We now have the general form of the solution within the layer and the two half-spaces on either side of the layer. Boundary conditions are needed to find the integration constants A, B, C, and D. For this system both U and dU/dz are continuous at $z = 0$ and $z = H$.

Problem d Use the results of problem b and problem c to show that these requirements impose the following constraints on the integration constants:

$$\left.\begin{aligned} A - C &= 0, \\ ik_0 A + k_1 D &= 0, \\ -Be^{ik_0 H} + C\cos k_1 H + D\sin k_1 H &= 0, \\ ik_0 Be^{ik_0 H} + k_1 C\cos k_1 H - k_1 D\sin k_1 H &= 0. \end{aligned}\right\} \tag{20.81}$$

This is a linear system of four equations for the four unknowns A, B, C, and D. Note that this is a homogeneous system of equations, because the right-hand sides vanish. Such a homogeneous system of equations only has nonzero solutions when the determinant of the system of equations vanishes.

Problem e Show that this requirement leads to the following condition:

$$\tan(k_1 H) = \frac{-2ik_0k_1}{k_1^2 + k_0^2}. \tag{20.82}$$

This equation is implicitly an equation for the wave-number k, because according to (20.78) and (20.79) both k_0 and k_1 are functions of the wave-number k. Equation (20.82) implies that the system can only support waves when the wavenumber k is such that expression (20.82) is satisfied. The system, strictly speaking, does not have normal modes, because the waves propagate in the x-direction. However, in the z-direction the waves only "fit" in the layer for very specific values of the wavenumber k. These waves are called "guided waves" because they propagate along the layer with a well-defined phase velocity that follows from the relation $c(\omega) = \omega/k$. Be careful not to confuse this phase velocity $c(\omega)$ with the velocities c_1 and c_0 in the layer and the half-spaces outside the layer. At this point we do not know yet what the phase velocities of the guided waves are.

The phase velocity follows from expression (20.82) because this expression is implicitly an equation for the wavenumber k. At this point we consider the case of a low-velocity layer, that is we assume that $c_1 < c_0$. In this case $1/c_0 < 1/c_1$. We look for guided waves with a wavenumber in the following interval: $\omega/c_0 < k < \omega/c_1$.

Problem f Show that in that case k_1 is real and k_0 is purely imaginary. Write $k_0 = i\kappa_0$ and show that

$$\kappa_0 = \sqrt{k^2 - \frac{\omega^2}{c_0^2}}. \tag{20.83}$$

Problem g Show that the solution decays exponentially away from the low-velocity channel both in the half-space $z < 0$ and the half-space $z > H$.

The fact that the waves decay exponentially with the distance to the low-velocity layer means that the guided waves are trapped near the low-velocity layer. Waves that decay exponentially are called *evanescent waves*.

Problem h Use (20.82) to show that the wavenumber of the guided waves satisfies the following relation:

$$\tan\sqrt{\frac{\omega^2}{c_1^2}-k^2}\,H=\frac{2\sqrt{k^2-\dfrac{\omega^2}{c_0^2}}\sqrt{\dfrac{\omega^2}{c_1^2}-k^2}}{\omega^2\left(\dfrac{1}{c_1^2}-\dfrac{1}{c_0^2}\right)}. \tag{20.84}$$

For a fixed value of ω this expression constitutes a constraint on the wavenumber k of the guided waves. Unfortunately, it is not possible to solve this equation for k in closed form. Such an equation is called a *transcendental equation.*

Problem j Make a sketch of both the left-hand side and the right-hand side of (20.84) as a function of k. Show that the two curves have a finite number of intersection points.

These intersection points correspond to the k-values of the guided waves. The corresponding phase velocity $c = \omega/k$ in general depends on the frequency ω. This means that these guided waves are *dispersive,* which means that the different frequency components travel with a different phase velocity. It is for this reason that (20.84) is called the *dispersion relation.*

Dispersive waves occur in many different situations. When electromagnetic waves propagate between plates or in a layered structure, guided waves result [53]. The atmosphere, and most importantly the ionosphere, is an excellent waveguide for electromagnetic waves [47]. This leads to a large variety of electromagnetic guided waves in the upper atmosphere with exotic names such as “pearls”, “whistlers”, “tweaks”, “hydromagnetic howling”, and “serpentine emissions”; colorful names associated with the sounds these phenomena would make if they were audible, or with the patterns they generate in frequency–time diagrams. These perturbations are excited for example by the electromagnetic fields generated by lightning. Guided waves play a crucial role in telecommunication, because light propagates through optical fibers as guided waves [61]. The fact that these waves are guided prohibits the light from propagating out of the fiber, and this allows for the transmission of light signals over extremely large distances.

In the Earth the wave velocity increases rapidly with depth. Elastic waves can be guided near the Earth’s surface and the different modes are called “Rayleigh waves” and “Love waves” [3]. These surface waves in the Earth are a prime tool for mapping the shear velocity within the Earth [98].

Since the surface waves in the Earth are trapped near the Earth’s surface, they effectively propagate in two dimensions rather than in three dimensions. The surface

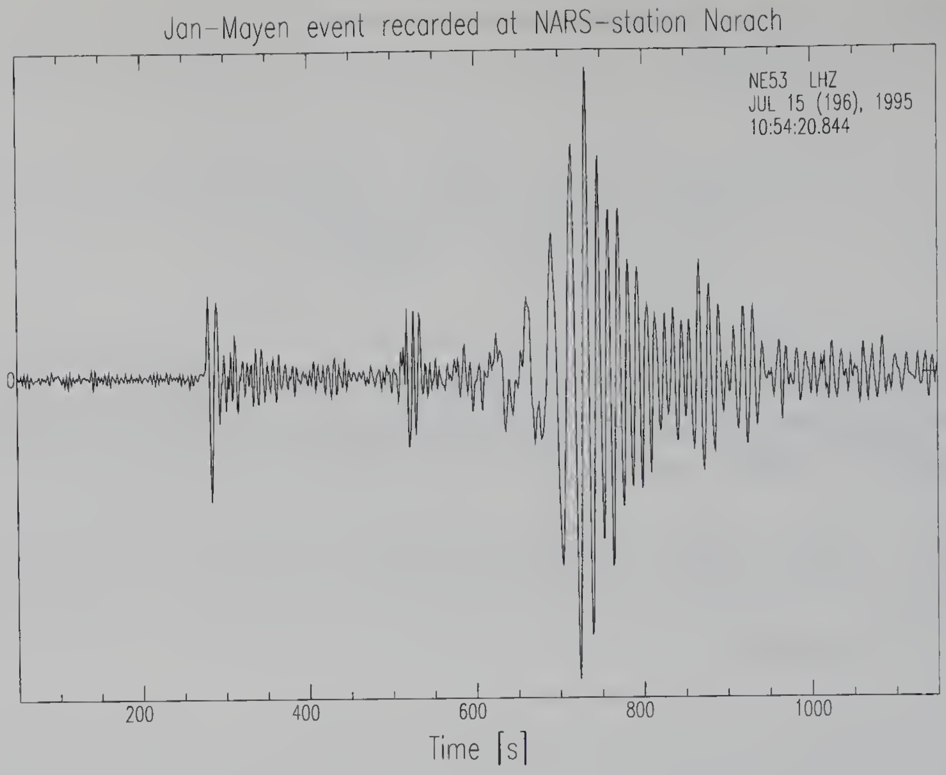

Fig. 20.11 Vertical component of the ground motion at a seismic station in Naroch (Belarus) after an earthquake at Jan-Mayen Island. This station is part of the Network of Autonomously Recording Seismographs (NARS) which is operated by Utrecht University.

waves therefore suffer less from geometrical spreading than the body waves that propagate through the interior of the Earth. For this reason, it is the surface waves that do most damage after an earthquake. This is illustrated in Figure 20.11 which shows the vertical displacement at a seismic station in Naroch (Belarus) after an earthquake at Jan-Mayen Island. Around $t = 300$ s and $t = 520$ s impulsive waves arrive: these are the body waves that travel through the interior of the Earth. The wave with the largest amplitude that arrives between $t = 650$ s and $t = 900$ s is the surface wave that is guided along the Earth's surface. Note that the waves that arrive around $t = 700$ s have a lower frequency content than the waves that arrive later, around $t = 850$ s. This is due to the fact that the group velocity of the low-frequency components of the surface wave is higher than the group velocity of the high-frequency components. Hence it is ultimately the dispersion of the Rayleigh waves that causes the change in the apparent frequency of the surface wave arrival.

20.9 Leaky modes

The guided waves in the previous section decay exponentially with the distance to the low-velocity layer. Intuitively, the fact that the waves are confined to a region near a low-velocity layer can be understood as follows. Waves are refracted from regions of high velocity to a region of low velocity. This means that the waves that stray out of the low-velocity channel are refracted back into the channel. Effectively this traps the waves in the vicinity of the channel. This explanation suggests that for a high-velocity channel the waves are refracted *away* from the channel. The resulting wave pattern then corresponds to waves that preferentially move away from the high-velocity layer. For this reason we consider in this section the waves that propagate through the system shown in Figure 20.10 but now we consider the case of a high-velocity layer where $c_1 > c_0$.

In this case, $1/c_1 < 1/c_0$, and we consider waves with a wavenumber that is confined to the following interval: $\omega/c_1 < k < \omega/c_0$.

Problem a Show that in this case the wavenumber k_1 is imaginary and that it can be written as $k_1 = i\kappa_1$, with

$$\kappa_1 = \sqrt{k^2 - \frac{\omega^2}{c_1^2}}, \tag{20.85}$$

and show that the dispersion relation (20.82) is given by:

$$\tan(i\kappa_1 H) = \frac{-2k_0\kappa_1}{\kappa_1^2 - k_0^2}. \tag{20.86}$$

Problem b Use the relation $\cos x = \left(e^{ix} + e^{-ix}\right)/2$ and the related expression for $\sin x$ to rewrite the dispersion relation (20.86) in the following form:

$$i\tanh(\kappa_1 H) = \frac{-2k_0\kappa_1}{\kappa_1^2 - k_0^2}. \tag{20.87}$$

In this expression all quantities are real when k is real. The factor i on the left-hand side implies that this equation cannot be satisfied for real values of k. The only way in which the dispersion relation (20.87) can be satisfied is if k is complex. What does it mean if the wavenumber is complex? Suppose that the dispersion relation is satisfied for a complex wavenumber $k = k_r + ik_i$, with k_r and k_i the real and imaginary parts. In the time domain a solution behaves for a fixed frequency as $U(k, z)e^{i(kx-\omega t)}$. This means that for complex values of the wavenumber the solution behaves as $U(k, z)e^{-k_i x}e^{i(k_r x-\omega t)}$. This is a wave that propagates in the x-direction with phase velocity $c = \omega/k_r$ and that decays exponentially with the propagation distance x.

The exponential decay of the wave with the propagation distance x is due to the fact that the wave energy refracts out of the high-velocity layer. A different way of understanding this exponential decay is to consider the character of the wave-field outside the layer.

Problem c Show that in the two half-spaces outside the high-velocity layer the waves propagate away from the layer. Hint: analyze the wavenumber k_0 in the half-spaces and consider the corresponding solution in these regions.

This means that wave energy is continuously radiated away from the high-velocity layer. The exponential decay of the mode with propagation distance x is thus due to the fact that wave energy continuously leaks out of the layer. For this reason one speaks of *leaky modes* [115]. In the Earth a well-observed leaky mode is the S-PL wave. This is a mode in which a transverse propagating wave in the mantle is coupled to a wave that is trapped in the Earth's crust.

In general there is no simple way to find the complex wavenumber k for which the dispersion relation (20.87) is satisfied. However, the presence of leaky modes can be seen in Figure 20.12 where the absolute value of the following function is shown in the complex k-plane:

$$F(k) \equiv 1/\left[i \tanh(\kappa_1 H) + \frac{2k_0\kappa_1}{\kappa_1^2 - k_0^2} \right]. \tag{20.88}$$

Problem d Show that this function is infinite for the k-values that correspond to a leaky mode.

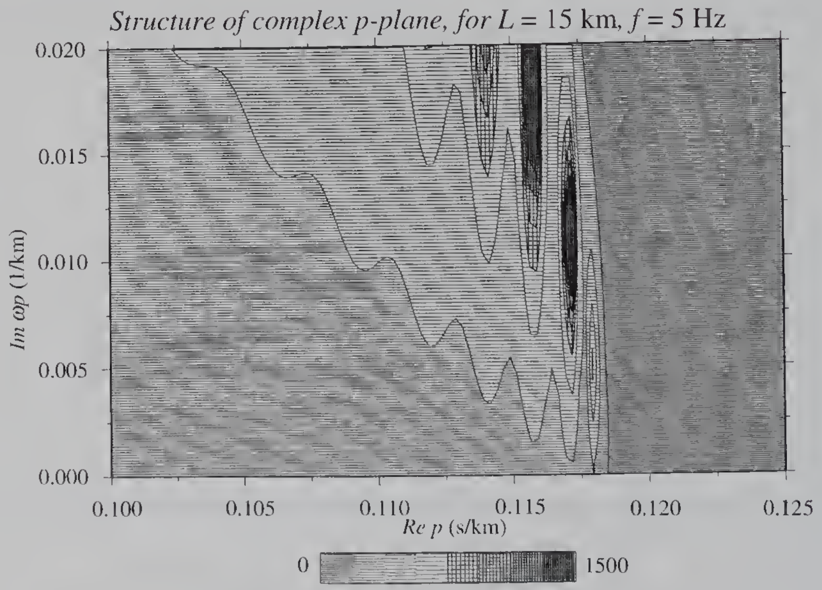

Fig. 20.12 Contour diagram of the function $|F(k)|$ for a high-velocity layer with velocity $c_1 = 8.4$ km/s and a thickness $H = 15$ km which is embedded between two half-spaces with velocity $c_0 = 8$ km/s, for waves with a frequency of 5 Hz. The horizontal axis is given by k_r/ω and the vertical axis by k_i.

The function $F(k)$ in Figure 20.12 is computed for a high-velocity layer with a thickness of 15 km and a velocity of 8.4 km/s that is embedded between two half-spaces with a velocity of 8 km/s. The frequency of the wave is 5 Hz. In this figure, the horizontal axis is given by $\Re e\ (p) = k_r/\omega$ while the vertical axis is given by $\Im m\,(\omega p) = k_i$. The quantity p is called the slowness and is defined as $p = k/\omega = 1/c(\omega)$. The leaky modes show up in Figure 20.12 as singularities of the function $F(k)$.

Problem e What is the propagation distance over which the amplitude of the mode with the lowest phase velocity decays with a factor $1/e$?

Leaky modes have been used by Gubbins and Snieder [46] to analyze waves which have propagated along a subduction zone. (A subduction zone is a plate in the Earth that slides downward in the mantle.) By a fortuitous geometry, compressive waves that are excited by earthquakes in the Tonga-Kermadec region travel to a seismic station in Wellington (New Zealand) and propagate for a large distance through the Tonga-Kermadec subduction zone. At the station in Wellington, a high-frequency wave arrives before the main compressive wave. This can be seen in Figure 20.13 where such a seismogram is shown band-pass filtered at

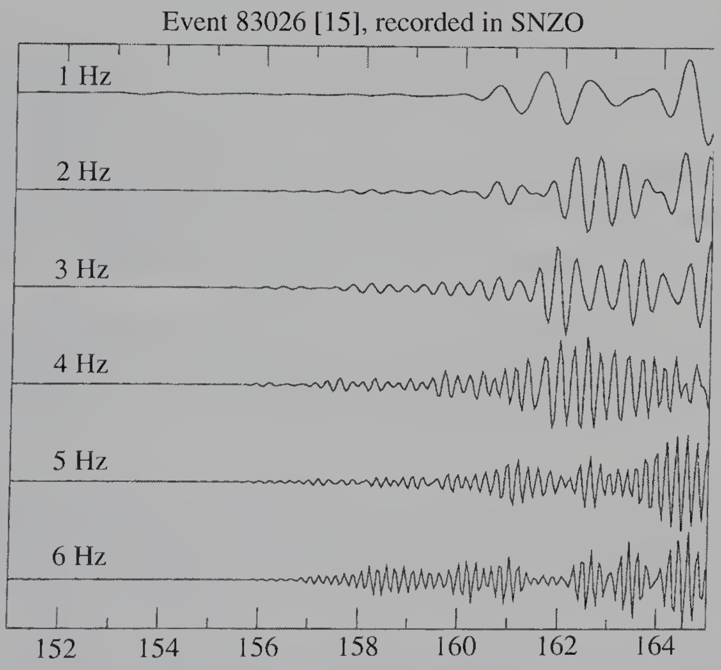

Fig. 20.13 Seismic waves recorded in Wellington after an earthquake in the Tonga-Kermadec subduction zone, from reference [46]. The different traces correspond to the waves band-pass filtered with a center frequency indicated at each trace. The horizontal axis gives the time since the earthquake occured in units of seconds.

different frequencies. It can clearly be seen that waves with a frequency around 6 Hz arrive before waves with a frequency around 1 Hz. This observation can be explained by the propagation of a leaky mode through the subduction zone. The physical reason why the high-frequency components arrive before the lower-frequency components is that the high-frequency waves "fit" in the high-velocity layer in the subducting plate, whereas the lower-frequency components do not fit in the high-velocity layer and are more influenced by the slower material outside the high-velocity layer. Of course, the energy leaks out of the high-velocity layer so that this arrival is very weak. From the data it could be inferred that in the subduction zone a high-velocity layer with a thickness between 6 and 10 km is present [46].

20.10 Radiation damping

Up to this point we have considered systems that are freely oscillating. When such systems are of finite extent, such a system displays undamped free oscillations. In the previous section leaky modes were introduced. In such a system, energy is radiated away, which leads to an exponential decay of waves that propagate through the system. In a similar way, a system that has normal modes when it is isolated from its surroundings can display damped oscillations when it is coupled to the external world.

As a simple prototype of such a system, consider a mass m that can move in the z-direction which is coupled to a spring with spring constant κ. The mass is attached to a string which is under a tension T and which has a mass ρ per unit length. The system is shown in Figure 20.14. The total force acting on the mass is the sum of the force $-\kappa z$ exerted by the spring and the force F_s that is generated by the string:

$$m\ddot{z} + \kappa z = F_s, \tag{20.89}$$

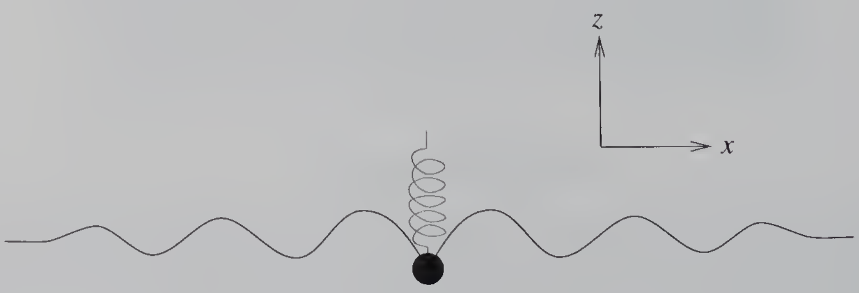

Fig. 20.14 Geometry of an oscillating mass that is coupled to a spring.

where z denotes the vertical displacement of the mass. The motion of the waves that propagate in the string is given by the wave equation:

$$u_{xx} - \frac{1}{c^2} u_{tt} = 0, \tag{20.90}$$

where u is the displacement of the string in the vertical direction and c is given by

$$c = \sqrt{\frac{T}{\rho}}. \tag{20.91}$$

Let us first consider the case in which no external force is present and the mass is not coupled to the spring.

Problem a Show that in that case the equation of motion is given by

$$\ddot{z} + \omega_0^2 z = 0, \tag{20.92}$$

with ω_0 given by

$$\omega_0 = \sqrt{\frac{\kappa}{m}}. \tag{20.93}$$

One can say that the mass that is not coupled to the string has one free oscillation with angular frequency ω_0. The fact that the system has one free oscillation is a consequence of the fact that this mass can move in the vertical direction only, hence it has only one degree of freedom.

Before we couple the mass to the string let us first analyze the wave motion in the string in the absence of the mass.

Problem b Show that any function $f(t - x/c)$ satisfies the wave equation (20.90). Show that this function describes a wave that moves in the positive x-direction with velocity c.

Problem c Show that any function $g(t + x/c)$ satisfies the wave equation (20.90) as well. Show that this function describes a wave that moves in the negative x-direction with velocity c.

The general solution is a superposition of the rightward and leftward moving waves:

$$u(x, t) = f\left(t - \frac{x}{c}\right) + g\left(t + \frac{x}{c}\right). \tag{20.94}$$

This general solution is called the *d'Alembert solution.*

Now we want to describe the motion of the coupled system. Let us first assume that the mass oscillates with a prescribed displacement $z(t)$ and find the waves that this displacement generates in the string. We consider here the situation in which there are no waves moving towards the mass. This means that to the right of the mass the waves can only move rightward and to the left of the mass there are only leftward moving waves.

Problem d Show that this radiation condition implies that the waves in the string are given by:

$$u(x,t) = \begin{cases} f\left(t - \frac{x}{c}\right) & \text{for} \quad x > 0 \\ g\left(t + \frac{x}{c}\right) & \text{for} \quad x < 0 \end{cases} . \tag{20.95}$$

Problem e At $x = 0$ the displacement of the mass is the same as the displacement of the string. Show that this implies that $f(t) = g(t) = z(t)$, so that

$$u(x,t) = \begin{cases} z\left(t - \frac{x}{c}\right) & \text{for} \quad x > 0 \\ z\left(t + \frac{x}{c}\right) & \text{for} \quad x < 0 \end{cases} . \tag{20.96}$$

Now we have solved the problem of finding the wave motion in the string given the motion of the mass. To complete our description of the system we also need to specify how the motion of the string affects the mass. In other words, we need to find the force F_s in (20.89) given the motion of the string. This force can be derived from Figure 20.15 . The vertical component F_+ of the force acting on the mass from the right-hand side of the string is given by $F_+ = T\sin\varphi$, where T is the tension in the string. When the motion in the spring is sufficiently weak we can approximate: $F_+ = T\sin\varphi \approx T\varphi \approx T\tan\varphi \approx Tu_x(x = 0^+, t)$. In the last identity we used that the derivative $u_x(x = 0^+, t)$ gives the slope of the string on the right of the point $x = 0$.

Problem f Use a similar reasoning to determine the force acting on the mass from the left-hand part of the spring and show that the net force acting on the spring

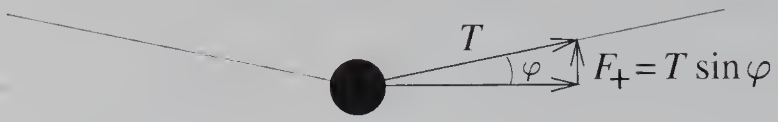

Fig. 20.15 Sketch of the force exerted by the spring on the mass.

is given by

$$F_s(t) = T\left(u_x(x = 0^+, t) - u_x(x = 0^-, t)\right), \tag{20.97}$$

where $u_x(x = 0^-, t)$ is the x-derivative of the displacement in the string just to the left of the mass.

Problem g Show that this expression implies that the net force that acts on the mass is equal to the *kink* in the spring at the location of the mass.

You may not feel comfortable with the fact the we used the approximation of a small angle φ in the derivation of (20.97). However, keep in mind that the wave equation (20.90) is derived using the same approximation and that this wave equation therefore is only valid for small displacements of the string.

At this point we have assembled all the ingredients for solving the coupled problem.

Problem h Use (20.96) and (20.97) to derive that the force exerted by the spring on the mass is given by

$$F_s(t) = -\frac{2T}{c}\,\dot{z}, \tag{20.98}$$

and that the motion of the mass is therefore given by:

$$\ddot{z} + \frac{2T}{mc}\dot{z} + \omega_0^2 z = 0. \tag{20.99}$$

It is interesting to compare this expression for the motion of the mass that is coupled to the string with the equation of motion (20.92) for the mass that is not coupled to the string. The string leads to a term $(2T/mc)\,\dot{z}$ in the equation of motion that damps the motion of the mass. How can we explain this damping physically? When the mass moves, the string moves with it at location $x = 0$. Any motion in the string at that point excites waves propagating in the string. This means that the string radiates wave energy away from the mass whenever the mass moves. Since energy is conserved, the energy that is radiated in the string must be withdrawn from the energy of the moving mass. This means that the mass loses energy whenever it moves; this effect is described by the damping term in equation (20.99). This damping process is called *radiation damping*, because it is the radiation of waves that damps the motion of the mass.

The system described in this section is extremely simple. However, it does contain the essential physics of radiation damping. Many systems in physics that display normal modes are not quite isolated from their surroundings. Interactions of the systems with their surroundings often lead to the radiation of energy, and hence to a damping of the oscillation of the system.

One example of such a system is an atom in an excited state. In the absence of external influences such an atom will not emit any light and will not decay. However, when such an atom can interact with electromagnetic fields, it can emit a photon and subsequently decay.

A second example is a charged particle that moves in a synchrotron. In the absence of external fields, such a particle will continue forever in a circular orbit without any change in its speed. In reality, a charged particle is coupled to electromagnetic fields. This has the effect that a charged particle that is accelerated emits electromagnetic radiation, called synchrotron radiation [53]. The radiated energy corresponds to an energy loss of the particle, so that the particle slows down. This is actually the reason why accelerators such as those used at CERN and Fermilab are so large. The acceleration of a particle in a circular orbit with radius r at the given velocity v is given by v^2/r. This means that for a fixed velocity v the larger is the radius of the orbit, the smaller is the acceleration, and the weaker is the energy loss due to the emission of synchrotron radiation. This is why one needs huge machines to accelerate tiny particles to an extreme energy.

Problem i The modes in the plate in Figure 20.1 are also damped because of radiation damping. What form of radiation is emitted by this oscillating plate?

21

Potential theory

Potential fields play an important role in physics and geophysics because they describe the behavior of gravitational and electric fields as well as a number of other fields. Conversely, measurements of potential fields provide important information about the internal structure of bodies. For example, measurements of the electric potential at the Earth's surface when a current is sent into the Earth give information about the electrical conductivity while measurements of the Earth's gravity field or geoid provide information about the mass distribution within the Earth.

An example of this can be seen in Figure 21.1 in which the gravity anomaly over the northern part of the Yucatan peninsula in Mexico is shown [49]. The coast is visible as a thin white line. Note the ring structure that is visible in the gravity signal. These rings have led to the discovery of the Chicxulub crater which was caused by the massive impact of a meteorite. Note that the diameter of the impact crater is about 150 km! This crater is presently hidden by thick layers of sediments: at the surface the only apparent imprint of this crater is the presence of underground water-filled caves called "cenotes" at the outer edge of the crater. It was the measurement of the gravity field that made it possible to find this massive impact crater.

The equation that the gravitational or electric potential satisfies depends critically on the Laplacian of the potential. As shown in Section 6.5 the gravitational field has the mass-density as its source:

$$(\nabla \cdot \mathbf{g}) = -4\pi G \rho. \tag{6.33}$$

The gravity field $\mathbf{g}$ is (minus) the gradient of the gravitational potential: $\mathbf{g} = -\nabla V$. This means that the gravitational potential satisfies the following partial differential equation:

$$\nabla^2 V(\mathbf{r}) = 4\pi G \rho. \tag{21.1}$$

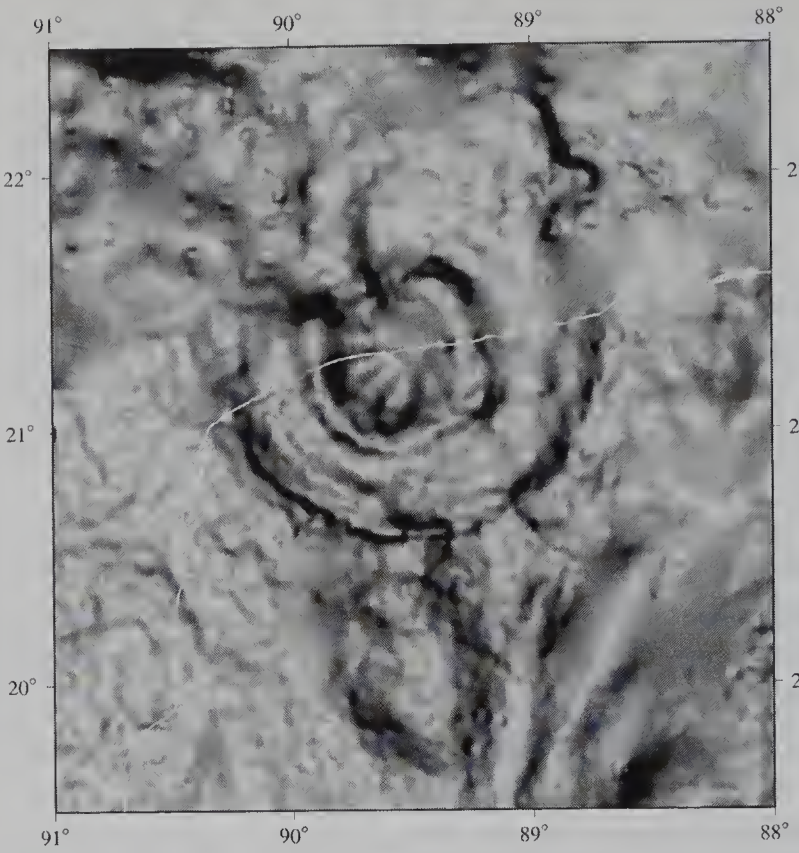

Fig. 21.1 Gravity field over the Chicxulub impact crater on the northern coast of Yucatan (Mexico). The coastline is shown by a white line. The numbers along the vertical and horizontal axes refer to the latitude and longitude respectively. The magnitude of the horizontal gradient of the Bouguer gravity anomaly is shown, details can be found in reference [49]. Courtesy of M. Pilkington and A. R. Hildebrand.

This equation is called Poisson's equation and is the prototype of the equations that occur in potential field theory. Note that the mathematical structures of the equations of the gravitational field and the electric field are identical (compare (6.23) and (6.33)), therefore the results derived in this chapter for the gravitational field can be used directly for the electric field as well by replacing the mass-density by the charge-density and by making the following replacement:

$$\underbrace{4\pi G}_{\text{gravity}} \Leftrightarrow \underbrace{-1/\varepsilon_0}_{\text{electrostatics}} \quad . \tag{21.2}$$

The theory of potential fields is treated in great detail by Blakeley [17].

21.1 Green's function of the gravitational potential

Poisson's equation (21.1) can be solved using a Green's function technique. In essence the derivation of the Green's function yields the well-known result that the gravitational potential for a point-mass m is given by $-Gm/r$. The use of Green's functions was introduced in great detail in Chapter 18. The Green's function $G(\mathbf{r}, \mathbf{r}')$ that describes the gravitational potential at location $\mathbf{r}$ generated by a point mass at location $\mathbf{r}'$ satisfies the following differential equation:

$$\nabla^2 G(\mathbf{r}, \mathbf{r}') = \delta\left(\mathbf{r} - \mathbf{r}'\right). \tag{21.3}$$

Take care not to confuse the Green's function $G(\mathbf{r}, \mathbf{r}')$ with the gravitational constant G.

Problem a Show that the solution of (21.1) is:

$$V(\mathbf{r}) = 4\pi G \int G(\mathbf{r}, \mathbf{r}')\rho(\mathbf{r}')\, dV'. \tag{21.4}$$

Problem b The differential equation (21.3) has translational invariance, and is invariant for rotations. Show that this implies that $G(\mathbf{r}, \mathbf{r}') = G(|\mathbf{r} - \mathbf{r}'|)$. Show by placing the point-mass at the origin by setting $\mathbf{r}' = 0$ that $G(r)$ satisfies

$$\nabla^2 G(r) = \delta(\mathbf{r}). \tag{21.5}$$

Problem c Use the expression for the Laplacian in spherical coordinates to show that for $r > 0$ (21.5) is given by

$$\frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2 \frac{\partial G(r)}{\partial r}\right) = 0. \tag{21.6}$$

Problem d Integrate this equation with respect to r to derive that the solution is given by $G(r) = A/r + B$, where A and B are integration constants.

The constant B in the potential does not contribute to the forces that are associated with this potential because $\nabla B = 0$. For this reason the arbitrary constant B can be taken to be equal to zero. The potential is therefore given by

$$G(r) = \frac{A}{r}. \tag{21.7}$$

Problem e The constant A can be found by integrating (21.5) over a sphere of radius R centered around the origin. Show that Gauss's theorem implies that $\int \nabla^2 G(r) dV = \oint \nabla G \cdot d\mathbf{S}$, use (21.7) in the right-hand side of this expression and show that this gives $A = -1/4\pi$. Note that this result is independent of the radius R that you have used.

Problem f Show that the Green's function is given by:

$$G(\mathbf{r}, \mathbf{r}') = -\frac{1}{4\pi}\frac{1}{|\mathbf{r}-\mathbf{r}'|}. \tag{21.8}$$

With (21.4) this implies that the gravitational potential is given by:

$$V(\mathbf{r}) = -G\int \frac{\rho(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|}dV'. \tag{21.9}$$

This general expression is useful for a variety of different purposes, and we will make extensive use of it. By taking the gradient of this expression one obtains the gravitational acceleration **g**. This acceleration was also derived in (8.5) for the special case of a spherically symmetric mass distribution. Surprisingly it is a nontrivial calculation to derive (8.5) by taking the gradient of (21.9).

21.2 Upward continuation in a flat geometry

Suppose that one has a body with variable mass in two dimensions and that the mass-density is only nonzero in the half-space $z < 0$. In this section we determine the gravitational potential V above the half-space when the potential is specified at the plane $z = 0$ that forms the upper boundary of this body. The geometry of this problem is sketched in Figure 21.2. This problem is of relevance for the interpretation of gravity measurements taken above the Earth's surface using aircraft or satellites because the first step in this interpretation is to relate the values of the potential at the Earth's surface to the measurements taken above the surface. This process is called *upward continuation.*

Mathematically the problem can be stated this way. Suppose one is given the function $V(x, z = 0)$, what is the function $V(x, z)$? When we know that there is no

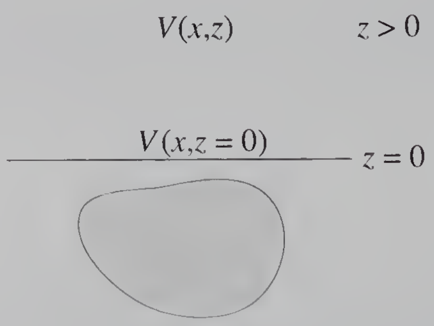

Fig. 21.2 Geometry of the upward continuation problem. A mass anomaly (shaded) leaves an imprint on the potential at $z = 0$. The upward continuation problem states how the potential at the surface $z = 0$ is related to the potential $V(x, z)$ at greater height.

mass above the surface it follows from (21.1) that the potential satisfies:

$$\nabla^2 V(\mathbf{r}) = 0 \qquad \text{for} \quad z > 0. \tag{21.10}$$

It is instructive to solve this problem by making a Fourier expansion of the potential in the variable x:

$$V(x, z) = \int_{-\infty}^{\infty} v(k, z) e^{ikx} dk. \tag{21.11}$$

Problem a Show that for $z = 0$ the Fourier coefficients can be expressed in terms of the known value of the potential at the edge of the half-space:

$$v(k, z = 0) = \frac{1}{2\pi} \int_{-\infty}^{\infty} V(x, z = 0) e^{-ikx} dx. \tag{21.12}$$

Problem b Use Poisson's equation (21.10) and the Fourier expansion (21.11) to derive that the Fourier components of the potential satisfy for $z > 0$ the following differential equation:

$$\frac{\partial^2 v(k, z)}{\partial z^2} - k^2 v(k, z) = 0. \tag{21.13}$$

Problem c Show that the general solution of this differential equation can be written as $v(k, z) = A(k)e^{+|k|z} + B(k)e^{-|k|z}$. The wave number k can be either positive or negative. By using the absolute value in the exponents it is explicit that the solution consists of a superposition of an exponentially growing solution (with z), and an exponentially decaying solution.

Since the potential must remain finite at great height ($z \to \infty$) the coefficient $A(k)$ must be equal to zero. Setting $z = 0$ shows that $B(k) = v(k, z = 0)$, so that the potential is given by:

$$V(x, z) = \int_{-\infty}^{\infty} v(k, z = 0) e^{ikx} e^{-|k|z} dk. \tag{21.14}$$

This expression is interesting because it states that the different Fourier components of the potential decay as $e^{-|k|z}$ with height.

Problem d Explain that this implies that the short-wavelength components in the potential field decay faster with height than the long-wavelength components.

The decrease of the Fourier components with the distance z to the surface is problematic when one wants to infer the mass-density in the body from measurements of the potential or from gravity at a great height above the surface, because the influence of mass perturbations on the gravitational field decays rapidly with height. The measurement of the short-wavelength component of the potential at a great height therefore carries virtually no information about the small-scale details of the density distribution within the Earth. This is the reason why gravity measurements from space are preferably carried out using satellites in low orbits rather than in high orbits. Similarly, for gravity surveys at sea, a gravity meter has been developed that is towed far below the sea surface [127]. The idea is that by towing the gravity meter closer to the sea bed, the gravity signal generated at the sub-surface for short wavelengths suffers less from the exponential decay due to upward continuation.

Problem e Take the gradient of (21.14) to find the vertical component of the gravity field. Use the resulting expression to show that the gravity field **g** is less sensitive to the exponential decay due to upward continuation than the potential V.

This last result is the reason why satellites in low orbits are used to measure the Earth's gravitational potential and satellites in high orbits are used to measure gravity. In fact, a space-borne gradiometer [89] is presently being developed. This instrument measures the *gradient* of the gravity vector by monitoring the differential motion between two masses in the satellite. Taking the gradient of the gravity leads to another factor of k in the Fourier expansion so that the effects of upward continuation are further reduced.

We now explicitly express the potential at height z to the potential at the surface $z = 0$.

Problem f Insert (21.12) into (21.14) to show that the upward continuation of the potential is given by:

$$V(x, z) = \int_{-\infty}^{\infty} H(x - x', z) V(x', z = 0)\, dx', \tag{21.15}$$

with

$$H(x, z) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-|k|z} e^{ikx}\, dk. \tag{21.16}$$

Note that (21.15) has exactly the same structure as (15.57) for a time-independent linear filter. The only difference is that the variable x now plays the role of the variable t in (15.57). This means that we can consider upward continuation as a linear filtering operation. The convolutional filter $H(x, z)$ maps the potential from the surface $z = 0$ onto the potential at height z.

Problem g Carry out the integral in expression (21.16) to show that this filter is given by:

$$H(x, z) = \frac{1}{\pi} \frac{z}{z^2 + x^2}. \tag{21.17}$$

Problem h Sketch this filter as a function of x for a large value of z and a small value of z.

Problem i Equation (21.15) implies that at the surface $z = 0$ this filter is given by $H(x, z = 0) = \delta(x)$, with $\delta(x)$ the Dirac delta function. Convince yourself of this by showing that $H(x, z)$ becomes more and more peaked round $x = 0$ when $z \to 0$ and by proving that for all values of z the filter function satisfies $\int_{-\infty}^{\infty} H(x, z)\, dx = 1$.

21.3 Upward continuation in a flat geometry in three dimensions

The analysis in the previous section is valid for a flat geometry in two dimensions. However, the theory can readily be extended to three dimensions by including another horizontal coordinate y in the derivation.

Problem a Show that the theory in the previous section up to equation (21.16) can be generalized by carrying out a Fourier transformation over both x and y. Show in particular that in three dimensions:

$$V(x, y, z) = \iint_{-\infty}^{\infty} H^{3D}(x - x', y - y', z) V(x', y', z = 0)\, dx' dy', \tag{21.18}$$

with

$$H^{3D}(x, y, z) = \frac{1}{(2\pi)^2} \iint_{-\infty}^{\infty} e^{-\sqrt{k_x^2 + k_y^2} z} e^{i(k_x x + k_y y)}\, dk_x dk_y. \tag{21.19}$$

The only difference with the case in two dimensions is that integral (21.19) leads to a different upward continuation filter than integral (21.16) for the two-dimensional case. The integral can be solved by switching the k-integral to cylindrical coordinates. The product $k_x x + k_y y$ can be written as $kr \cos\varphi$, where

k and r are the length of the **k**-vector and the position vector in the horizontal plane.

Problem b Use this to show that H^{3D} can be written as

$$H^{3D}(x, y, z) = \frac{1}{(2\pi)^2} \int_0^\infty \int_0^{2\pi} ke^{-kz} e^{ikr\cos\varphi} d\varphi dk. \tag{21.20}$$

Note the factor k in the integrand.

Problem c As shown in Section 24.5, the Bessel function has the following integral representation:

$$J_0(x) = \frac{1}{2\pi} \int_0^{2\pi} e^{ix\cos\theta} d\theta. \tag{21.21}$$

Use this result to write the upward continuation filter as

$$H^{3D}(x, y, z) = \frac{1}{2\pi} \int_0^\infty e^{-kz} J_0(kr) k dk. \tag{21.22}$$

It appears that we have only made the problem more complex, because the integral of the Bessel function is not trivial. Fortunately, books and tables exist with a bewildering collection of integrals. For example, in equation (6.621.1) of Gradshteyn and Ryzhik [45] you can find an expression for the following integral: $\int_0^\infty e^{-\alpha x} J_\nu(\beta x) x^{\mu-1} dx$.

Problem d What are the values of α, ν, β, and μ if we want to use this integral to solve the integration in (21.22)?

Take a look at the form of the integral in Gradshteyn and Ryzhik [45]. You will probably be discouraged by what you find because the result is expressed in hypergeometric functions, which means that you now have the new problem of finding out what these functions are. There is, however, a way out because (6.611.1) of reference [45] gives the following integral:

$$\int_0^\infty e^{-\alpha x} J_\nu(\beta x) dx = \frac{\beta^{-\nu} \left\{ \sqrt{\alpha^2 + \beta^2} - \alpha \right\}^\nu}{\sqrt{\alpha^2 + \beta^2}}. \tag{21.23}$$

This is not quite the integral that we want because it does not contain a term that corresponds to the factor k in (21.22). However, we can introduce such a factor by differentiating (21.23) with respect to α.

Problem e Do this to show that the upward continuation operator is given by

$$H^{3D}(x, y, z) = \frac{1}{2\pi} \frac{z}{\left(x^2 + y^2 + z^2\right)^{3/2}}. \tag{21.24}$$

Hint: you can make the problem simpler by first inserting the appropriate value of ν in (21.23).

Problem f Compare the upward continuation operator (21.24) for three dimensions with the corresponding operator for two dimensions in (21.17). Which of these operators decays more rapidly as a function of the horizontal distance? Can you explain this difference physically?

Problem g In Section 21.2 you showed that the integral of the upward continuation operator over the horizontal distance is equal to 1. Show that the same holds in three dimensions, that is show that $\int\int_{-\infty}^{\infty} H^{3D}(x, y, z)dxdy = 1$. The integration simplifies by using cylindrical coordinates.

Comparing the upward continuation operators in different dimensions one finds that these operators are different functions of the horizontal distance in the space domain. However, a comparison of (21.16) with (21.19) shows that in the wavenumber domain the upward continuation operators in two and three dimensions have the same dependence on wave-number. The same is actually true for the Green's function of the wave equation in one, two, or three dimensions. As you can see in Section 19.4 these Green's functions are very different in the space domain, but one can show that in the wavenumber domain they are given in each dimension by the same expression.

21.4 Gravity field of the Earth

In this section we obtain an expression for the gravitational potential outside the Earth for an arbitrary distribution of the mass density $\rho(\mathbf{r})$ within the Earth. This could be done be using the Green's function that is appropriate for Poisson's equation (21.1). As an alternative we solve the problem here by expanding both the mass-density and the potential in spherical harmonics and by using a Green's function technique for every component in the spherical harmonics expansion separately.

When using spherical harmonics, the natural coordinate system is a system of spherical coordinates. For every value of the radius r both the density and the potential can be expanded in spherical harmonics:

$$\rho(r,\theta,\varphi)=\sum_{l=0}^{\infty}\sum_{m=-l}^{l}\rho_{lm}(r)Y_{lm}(\theta,\varphi) \tag{21.25}$$

and

$$V(r,\theta,\varphi)=\sum_{l=0}^{\infty}\sum_{m=-l}^{l}V_{lm}(r)Y_{lm}(\theta,\varphi). \tag{21.26}$$

Problem a Use the orthogonality relation (20.50) of spherical harmonics to show that the expansion coefficients for the density are given by:

$$\rho_{lm}(r)=\int Y_{lm}^{*}(\theta,\varphi)\rho(r,\theta,\varphi)\,d\Omega, \tag{21.27}$$

where $\int(\cdots)\,d\Omega$ denotes an integration over the unit sphere.

Equation (21.1) for the gravitational potential contains the Laplacian. The Laplacian in spherical coordinates can be decomposed as

$$\nabla^2=\frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2\frac{\partial}{\partial r}\right)+\frac{1}{r^2}\nabla_1^2, \tag{21.28}$$

with ∇_1^2 the Laplacian on the unit sphere:

$$\nabla_1^2=\frac{1}{\sin\theta}\frac{\partial}{\partial\theta}\left(\sin\theta\frac{\partial}{\partial\theta}\right)+\frac{1}{\sin^2\theta}\frac{\partial^2}{\partial\varphi^2}. \tag{20.34}$$

The reason why an expansion in spherical harmonics is used for the density and the potential is that the spherical harmonics are the eigenfunctions of the operator ∇_1^2 (see (20.33) or p. 379 of Butkov [24]):

$$\nabla_1^2 Y_{lm}=-l\,(l+1)\,Y_{lm}. \tag{20.33}$$

Problem b Insert (21.25) and (21.26) into the Laplace equation, and use (20.33) and (20.34) for the Laplacian of the spherical harmonics to show that the expansion coefficients $V_{lm}(r)$ of the potential satisfy the following differential equation:

$$\frac{1}{r^2}\frac{\partial}{\partial r}\left[r^2\frac{\partial V_{lm}(r)}{\partial r}\right]-\frac{l\,(l+1)}{r^2}V_{lm}(r)=4\pi G\rho_{lm}(r). \tag{21.29}$$

What we have gained by making the expansion in spherical harmonics is that (21.29) is an ordinary differential equation in the variable r whereas the original equation (21.1) is a partial differential equation in the variables r, θ and φ. The differential equation (21.29) can be solved using the Green's function technique described in Section 18.3. Let us first consider a mass $\delta(r - r')$ located at a radius r'. The response to this mass is the Green's function G_l that satisfies the following differential equation:

$$\frac{1}{r^2}\frac{\partial}{\partial r}\left[r^2\frac{\partial G_l(r,r')}{\partial r}\right] - \frac{l\,(l+1)}{r^2}G_l(r,r') = \delta\left(r - r'\right). \tag{21.30}$$

Note that this equation depends on the angular order l but not on the angular degree m. For this reason the Green's function $G_l(r, r')$ depends on l but not on m.

Problem c The Green's function can be found by first solving the differential equation (21.30) for $r \neq r'$. Show that when $r \neq r'$ the general solution of the differential equation (21.30) can be written as $G_l = Ar^l + Br^{-(l+1)}$, where the constants A and B do not depend on r.

Problem d In the regions $r < r'$ and $r > r'$ the constants A and B in general have different values. Show that the requirement that the potential is everywhere finite implies that $B = 0$ for $r < r'$ and that $A = 0$ for $r > r'$. The solution can therefore be written as:

$$G_l(r,r') = \begin{cases} Ar^l & \text{for} \quad r < r' \\ Br^{-(l+1)} & \text{for} \quad r > r' \end{cases}. \tag{21.31}$$

The integration constants follow in the same way as in the analysis of Section 18.3. One constraint on the integration constants follows from the requirement that the Green's function is continuous in the point $r = r'$. The other constraint follows by multiplying (21.30) by r^2 and integrating the resulting equation over r from $r' - \varepsilon$ to $r' + \varepsilon$.

Problem e Show by taking the limit $\varepsilon \to 0$ that this leads to the requirement

$$\left[r^2\frac{\partial G_l(r,r')}{\partial r}\right]_{r=r'-\varepsilon}^{r=r'+\varepsilon} = r'^2. \tag{21.32}$$

Problem f Use this condition with the continuity of G_l to find the coefficients A and B and show that the Green's function is given by:

$$G_l(r, r') = \begin{cases} -\dfrac{1}{(2l+1)} \dfrac{r^l}{r'^{(l-1)}} & \text{for} \quad r < r' \\ -\dfrac{1}{(2l+1)} \dfrac{r'^{(l+2)}}{r^{(l+1)}} & \text{for} \quad r > r' \end{cases} . \tag{21.33}$$

Problem g Use this result to derive that the solution of (21.29) is:

$$\begin{aligned} V_{lm}(r) = &-\frac{4\pi G}{(2l+1)} \frac{1}{r^{l+1}} \int_0^r \rho_{lm}(r') r'^{(l+2)} dr' \\ &-\frac{4\pi G}{(2l+1)} r^l \int_r^\infty \rho_{lm}(r') \frac{1}{r'^{(l-2)}} dr'. \end{aligned} \tag{21.34}$$

Hint: split the integration over r' into the interval $0 < r' < r$ and the interval $r' > r$.

Problem h Let us now consider the potential outside the Earth. The radius of the Earth is denoted by the symbol a. Use the above expression to show that the potential outside the Earth is given by

$$V(r, \theta, \varphi) = -\sum_{l=0}^{\infty} \sum_{m=-l}^{l} \frac{4\pi G}{(2l+1)} \frac{1}{r^{l+1}} \int_0^a \rho_{lm}(r') r'^{(l+2)} dr' \; Y_{lm}(\theta, \varphi). \tag{21.35}$$

Problem i Eliminate ρ_{lm} using the result of problem a and show that the potential is finally given by:

$$\begin{aligned} V(r, \theta, \varphi) = &-\sum_{l=0}^{\infty} \sum_{m=-l}^{l} \frac{4\pi G}{(2l+1)} \frac{1}{r^{l+1}} \\ &\times \int_0^a \rho(r', \theta', \varphi') r'^l Y_{lm}^*(\theta', \varphi') \, dV' \; Y_{lm}(\theta, \varphi). \end{aligned} \tag{21.36}$$

Note that the integration in this expression is over the volume of the Earth rather than over the distance r' to the Earth's center.

Let us reflect on the relation between this result and the derivation of upward continuation in a Cartesian geometry of Section 21.2. Equation (21.36) can be compared with (21.14). In (21.14) the potential is written as an integration over wavenumber and the potential is expanded in basis functions e^{ikx}, whereas in (21.36) the potential is written as a summation over the degree l and order m and the potential is expanded in basis functions Y_{lm}. In both expressions the potential is written as a sum over basis functions with increasingly shorter wavelength as the summation index l or the integration variable k increases. The decay of the potential with height is in both geometries faster for a potential with rapid horizontal variations than for a potential with smooth horizontal variations. In both cases the potential decreases when the height z (or the radius r) increases. In a flat geometry the potential decreases as $e^{-|k|z}$ whereas in a spherical geometry the potential decreases as $r^{-(l+1)}$. This difference in the reduction of the potential with distance is due to the difference in the geometry in the two problems.

Expressions (21.9) and (21.36) both express the gravitational potential due to the same density distribution $\rho(\mathbf{r})$, therefore these expressions must be identical.

Problem j Use the equivalence of these expressions to derive that for $r' < r$ the following identity holds:

$$\frac{1}{|\mathbf{r}-\mathbf{r}'|} = \sum_{l=0}^{\infty}\sum_{m=-l}^{l}\frac{4\pi}{(2l+1)}Y_{lm}(\theta,\varphi)Y_{lm}^{*}(\theta',\varphi')\frac{r'^{l}}{r^{l+1}}. \tag{21.37}$$

The derivation in this section could also have been made using (21.37) as a starting point because this expression can be derived by using the *generating function* of Legendre polynomials and by using the addition theorem to obtain the m-summation [5, 53]. However, these concepts are not needed in the treatment in this section which is based only on the expansion of functions in spherical harmonics and on the use of Green's functions.

As a last exercise let us consider the special case of a spherically symmetric mass distribution: $\rho = \rho(r)$. For such a mass distribution the potential is given by

$$V(\mathbf{r}) = -\frac{GM}{r}, \tag{21.38}$$

where M is the total mass of the body. The gradient of this potential is indeed equal to the gravitational acceleration given in (8.5) for a spherically symmetric mass M.

Problem i Derive the potential (21.38) from (21.36) by considering the special case that the mass-density depends only on the radius.

21.5 Dipoles, quadrupoles, and general relativity

We have seen in Section 8.2 that a spherically symmetric mass leads to a gravitational field $\mathbf{g}(\mathbf{r}) = -GM\hat{\mathbf{r}}/r^2$, which corresponds to a gravitational potential $V(\mathbf{r}) = -GM/r$. Similarly, the electric potential due to a spherically symmetric charge distribution is given by $V(\mathbf{r}) = q/4\pi\varepsilon_0 r$, where q is the total charge. In this section we investigate what happens if we place a positive charge and a negative charge close together. Since there is no negative mass, we treat for the moment the electric potential, but we will see in Section 21.6 that the results also have a bearing on the gravitational potential. The theory developed here is not only important in electrostatics, it also accounts for the measurable effect of the ellipsoidal shape of the Earth on its gravitational field. This application is treated in more detail in Section 21.7.

Consider a charge distribution that consists of a positive charge $+q$ placed at position $\mathbf{a}/2$ and a negative charge $-q$ placed at position $-\mathbf{a}/2$ as shown in Figure 21.3.

Problem a The total charge of this system is zero. What would you expect the electric potential to be at positions that are very far from the charges compared to the distance between the charges?

Problem b The potential follows by adding the potentials for the two point charges. Show that the electric potential generated by these two charges is given by

$$4\pi\varepsilon_0 V(\mathbf{r}) = \frac{q}{|\mathbf{r}-\mathbf{a}/2|} - \frac{q}{|\mathbf{r}+\mathbf{a}/2|}. \tag{21.39}$$

Problem c Ultimately we will place the charges very close to the origin by taking the limit $a \rightarrow 0$. We can therefore restrict our attention to the special case that $a \ll r$. Use a first order Taylor expansion to show that up to order a:

$$\frac{1}{|\mathbf{r}-\mathbf{a}/2|} = \frac{1}{r} + \frac{1}{2r^3}(\mathbf{r}\cdot\mathbf{a}). \tag{21.40}$$

Hint: use that $|\mathbf{r}-\mathbf{a}/2|^{-1} = [(\mathbf{r}-\mathbf{a}/2)\cdot(\mathbf{r}-\mathbf{a}/2)]^{-1/2}$.

Problem d Insert this into (21.39) and derive that the electric potential is given by:

$$4\pi\varepsilon_0 V(\mathbf{r}) = -\frac{q\,(\mathbf{r}\cdot\mathbf{a})}{r^3}. \tag{21.41}$$

Now suppose we bring the charges in Figure 21.3 closer and closer together, and suppose we let the charge q increase so that the product $\mathbf{p} = q\mathbf{a}$ is constant, then the electric potential is given by:

$$4\pi\varepsilon_0 V(\mathbf{r}) = -\frac{(\hat{\mathbf{r}}\cdot\mathbf{p})}{r^2}, \tag{21.42}$$

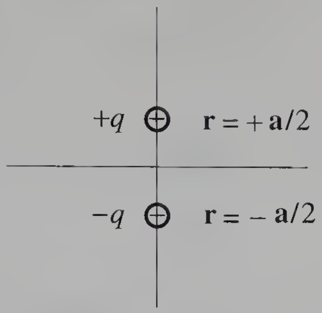

Fig. 21.3 Two opposite charges that constitute an electric dipole.

where we have used that $\mathbf{r} = r\hat{\mathbf{r}}$. The vector $\mathbf{p}$ is called the *dipole vector*. We will see in the next section how the dipole vector can be defined for arbitrary charge or mass distributions.

In problem a you might have guessed that the electric potential would go to zero at great distance. Of course the potential due to the combined charges goes to zero much faster than the potential due to a single charge only: the electric potential of a dipole vanishes as $1/r^2$ compared to the $1/r$ decay of the potential for a single charge. Many physical systems, such as neutral atoms, consist of neutral combinations of positive and negative charges. The lesson we learn from (21.41) is that such a neutral combination of charges may generate a nonzero electric field and that such a system will in general interact with other electromagnetic systems. For example, atoms interact to leading order with the radiation field (light) through their *dipole moment* [91]. In chemistry, the dipole moment of molecules plays a crucial role in the distinction between polar and apolar substances. Water would not have its many wonderful properties if it did not have a dipole moment.

Let us now consider the electric field generated by an electric dipole.

Problem e Take the gradient of (21.42) to show that this field is given by

$$\mathbf{E}(\mathbf{r}) = \frac{1}{4\pi\varepsilon_0 r^3}\left[\mathbf{p} - 3\hat{\mathbf{r}}\,(\hat{\mathbf{r}}\cdot\mathbf{p})\right] . \tag{21.43}$$

Hint: either use the expression of the gradient in spherical coordinates or take the gradient in Cartesian coordinates and use (6.11).

The electric field generated by an electric dipole has the same form as the magnetic field generated by a magnetic dipole as shown in (6.5). The mathematical reason for this is that the magnetic field satisfies (6.28) which states that $(\nabla\cdot\mathbf{B}) = 0$, while the electric field in free space satisfies according to (6.23) the field equation: $(\nabla\cdot\mathbf{E}) = 0$. However, there is an important difference. The electric field is generated by electric charges, and this field satisfies the equation $(\nabla\cdot\mathbf{E}) = \rho(\mathbf{r})/\varepsilon_0$. In

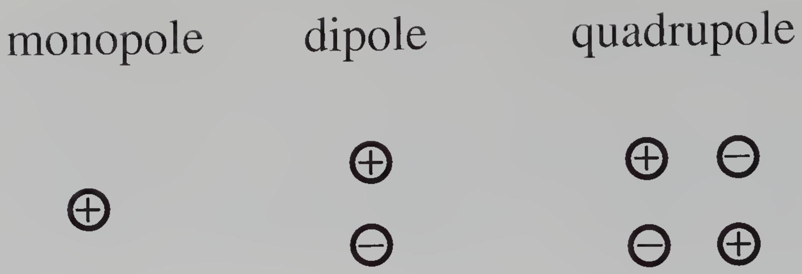

Fig. 21.4 The definition of the monopole, dipole, and quadrupole in terms of electric charges.

the example in this section we created a dipole field by taking two opposite charges and putting them closer and closer together. However, the magnetic field satisfies $(\nabla \cdot \mathbf{B}) = 0$ *everywhere*. The reason for this is that the magnetic equivalent of the electric charge, the *magnetic monopole*, has not been discovered in nature.

The fact that magnetic monopoles have not been observed in nature seems puzzling, because we have seen that the magnetic dipole field has the same form as the electric dipole field which was constructed by putting two opposite electric charges close together. The reason for the analogy between an electric and a magnetic dipole field is not that the magnetic dipole can be seen as a combination of positive and negative magnetic "charges" placed close together. In the context of classical electromagnetism, the magnetic dipole field is generated by a current that runs in a small circular loop. On a microscopic scale a magnetic dipole is generated by the *spin* of particles. This means that the electric and magnetic dipole fields have fundamentally different origins.

The starting point of the derivation in this section is the electric field of a single point charge. Such a single point charge is called a monopole, see Figure 21.4. The field of this charge decays as $1/r^2$. If we put two opposite charges close together we can create a dipole, see Figure 21.4. Its electric field is derived in (21.43); this field decays as $1/r^3$. We can also put two opposite dipoles together as shown in Figure 21.4. The resulting charge distribution is called a quadrupole. To leading order the electric fields of the dipoles that constitute the quadrupole cancel, and we will see in the next section that the electric potential for a quadrupole decays as $1/r^3$ so that the electric field decays with distance as $1/r^4$.

You may wonder whether the concept of a dipole or quadrupole can also be used for the gravity field because for the electric field these concepts are based on the presence of both positive and negative charges whereas we know that only positive mass occurs in nature. However, there is nothing to keep us from computing the gravitational field for a negative mass, and this is actually quite useful. As an example, let us consider a double star that consists of two heavy stars which rotate around their joint center of gravity. The first-order field is the monopole field

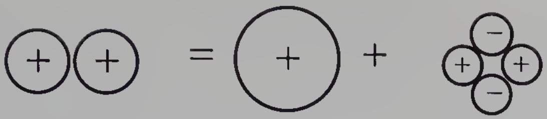

Fig. 21.5 The decomposition of a double star in a gravitational monopole and a gravitational quadrupole.

that is generated by the joint mass of the stars. However, as shown in Figure 21.5 the mass of the two stars can be seen as approximately the sum of a monopole and a quadrupole consisting of two positive and two negative masses. Since the stars rotate, the gravitational quadrupole rotates as well, and this is the reason why rotating double stars are seen as a prime source for the generation of *gravitational waves* [76]. However, gravitational waves that spread in space with time cannot be described by the classic expression (21.1) for the gravitational potential.

Problem g Can you explain why (21.1) cannot account for propagating gravitational waves?

A proper description of gravitational waves depends on the general theory of relativity [76]. Huge detectors for gravitational waves are being developed [8], because these waves can be used to investigate the theory of general relativity as well as the astronomical objects that generate gravitational waves.

21.6 Multipole expansion

Now that we have learned that the concepts of the monopole, dipole, and quadrupole are relevant for both the electric field and the gravity field we continue the analysis with the gravitational field. In this section we derive the *multipole expansion* in which the total field is written as a superposition of a monopole field, a dipole field, a quadrupole field, an octupole field, etc.

Consider the situation shown in Figure 21.6 in which a finite body has a mass-density $\rho(\mathbf{r}')$. The gravitational potential generated by this mass is given by:

$$V(\mathbf{r}) = -G \int \frac{\rho(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|} dV' . \tag{21.9}$$

We consider the potential at a distance that is much larger than the size of the body. Since the integration variable $\mathbf{r}'$ is limited by the size of the body, a "large distance" means in this context that $r \gg r'$. We therefore make a Taylor expansion of the term $1/\left|\mathbf{r}-\mathbf{r}'\right|$ in the parameter $\left(r'/r\right)$ which is much smaller than unity.

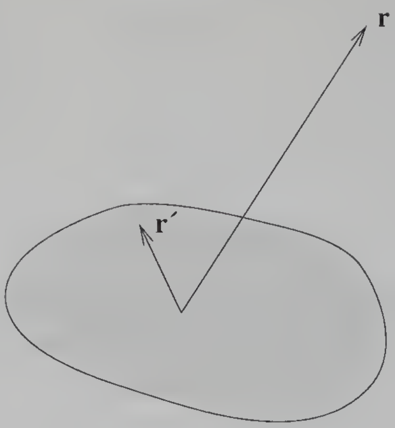

Fig. 21.6 Definition of the integration variable $\mathbf{r}'$ within the mass and the observation point $\mathbf{r}$ outside the mass.

Problem a Show that

$$\left|\mathbf{r}-\mathbf{r}'\right| = \sqrt{r^2 - 2\left(\mathbf{r}\cdot\mathbf{r}'\right) + r'^2}. \tag{21.44}$$

Problem b Use a Taylor expansion in the small parameter r'/r to show that:

$$\frac{1}{\left|\mathbf{r}-\mathbf{r}'\right|} = \frac{1}{r}\left\{1 + \frac{1}{r}\left(\hat{\mathbf{r}}\cdot\mathbf{r}'\right) + \frac{1}{2r^2}\left[3\left(\hat{\mathbf{r}}\cdot\mathbf{r}'\right)^2 - r'^2\right] + O\left(\left(\frac{r'}{r}\right)^3\right)\right\}. \tag{21.45}$$

Be careful that you account for all the terms of order r' correctly. Also be aware of the distinction between the position vector $\mathbf{r}$ and the unit vector $\hat{\mathbf{r}}$.

From this point on we will ignore the terms of order $\left(r'/r\right)^3$.

Problem c Insert the expansion (21.45) into (21.9) and show that the gravitational potential can be written as a sum of different contributions:

$$V(\mathbf{r}) = V_{mon}(\mathbf{r}) + V_{dip}(\mathbf{r}) + V_{qua}(\mathbf{r}) + \cdots, \tag{21.46}$$

with

$$V_{mon}(\mathbf{r}) = -\frac{G}{r}\int \rho(\mathbf{r}')dV', \tag{21.47}$$

$$V_{dip}(\mathbf{r}) = -\frac{G}{r^2}\int \rho(\mathbf{r}')\left(\hat{\mathbf{r}}\cdot\mathbf{r}'\right)dV', \tag{21.48}$$

$$V_{qua}(\mathbf{r}) = -\frac{G}{2r^3}\int \rho(\mathbf{r}')\left[3\left(\hat{\mathbf{r}}\cdot\mathbf{r}'\right)^2 - r'^2\right]dV'. \tag{21.49}$$

It thus follows that the gravitational potential can be written as the sum of terms that decay with increasing powers of r^{-n}. Let us analyze these terms in turn. The term $V_{mon}(\mathbf{r})$ in (21.47) is the simplest since the volume integral of the mass-density is simply the total mass of the body: $\int \rho(\mathbf{r}')dV' = M$. This means that this term is given by

$$V_{mon}(\mathbf{r}) = -\frac{GM}{r}. \tag{21.50}$$

This is the potential generated by a point mass M. To leading order, the gravitational field is the same as if all the mass of the body were concentrated in the origin. The mass distribution within the body does not affect this part of the gravitational field at all. Because the resulting field is the same as for a point mass, this field is called the *monopole field*.

For the analysis of the term $V_{dip}(\mathbf{r})$ in (21.48) it is useful to define the center of gravity $\mathbf{r}_g$ of the body:

$$\mathbf{r}_g \equiv \frac{\int \rho(\mathbf{r}')\mathbf{r}' dV'}{\int \rho(\mathbf{r}')\, dV'}. \tag{21.51}$$

This is simply a weighted average of the position vector with the mass-density as weight functions. Note that the word “weight” here has a double meaning!

Problem d Show that $V_{dip}(\mathbf{r})$ is given by:

$$V_{dip}(\mathbf{r}) = -\frac{GM}{r^2}\left(\hat{\mathbf{r}} \cdot \mathbf{r}_g\right). \tag{21.52}$$

Note that this potential has exactly the same form as the potential (21.42) for an electric dipole. For this reason $V_{dip}(\mathbf{r})$ is called the dipole field.

Problem e Compared to the monopole term, the dipole term decays as $1/r^2$ rather than $1/r$. The monopole term does not depend on $\hat{\mathbf{r}}$, the direction of observation. Show that the dipole term varies with the direction of observation as $\cos\theta$ and show how the angle θ must be defined.

Problem f You may be puzzled by the fact that the gravitational potential contains a dipole term, despite the fact that there is no negative mass. Draw a figure similar to Figure 21.5 to show that a displaced mass can be written as an undisplaced mass plus a mass dipole.

Of course, one is free in the choice of the origin of the coordinate system. If one chooses the origin to be at the center of mass of the body, then $\mathbf{r}_g = 0$ and the dipole term vanishes.

We now analyze the term $V_{qua}(\mathbf{r})$ in (21.49). It can be seen from this expression that this term decays with distance as $1/r^3$. For this reason, this term is called the quadrupole field. The dependence of the quadrupole field on the direction is more complex than for the monopole field and the dipole field. In the determination of the directional dependence of the quadrupole term it is useful to use the double contraction between two tensors. Tensors are treated in Chapter 22. For the moment you only need to know that the double contraction is defined as:

$$(\mathbf{A} : \mathbf{B}) \equiv \sum_{i,j} A_{ij} B_{ij} . \tag{21.53}$$

The double contraction generalizes the concept of the inner product of two vectors $(\mathbf{a} \cdot \mathbf{b}) = \sum_i a_i b_i$ to matrices or tensors of rank two. A double contraction occurs for example in the following identity

$$1 = (\hat{\mathbf{r}} \cdot \hat{\mathbf{r}}) = (\hat{\mathbf{r}} \cdot \mathbf{I}\hat{\mathbf{r}}) = (\hat{\mathbf{r}}\hat{\mathbf{r}} : \mathbf{I}) , \tag{21.54}$$

where $\mathbf{I}$ is the identity operator. Note that the term $\hat{\mathbf{r}}\hat{\mathbf{r}}$ is a dyad. If you are unfamiliar with the concept of a dyad you may want to look at Section 13.1 before continuing with this chapter.

Problem g Use these results to show that $V_{qua}(\mathbf{r})$ can be written as:

$$V_{qua}(\mathbf{r}) = -\frac{G}{2r^3} (\hat{\mathbf{r}}\hat{\mathbf{r}} : \mathbf{T}) , \tag{21.55}$$

where $\mathbf{T}$ is the *inertia tensor* defined as

$$\mathbf{T} = \int \rho(\mathbf{r}) \left(3\mathbf{r}\mathbf{r} - \mathbf{I}r^2 \right) dV . \tag{21.56}$$

Note that we have renamed the integration variable $\mathbf{r}'$ in the quadrupole moment tensor as $\mathbf{r}$.

Problem h Show that in explicit matrix notation $\mathbf{T}$ is given by:

$$\mathbf{T} = \int \rho(\mathbf{r}) \begin{pmatrix} 2x^2 - y^2 - z^2 & 3xy & 3xz \\ 3xy & 2y^2 - x^2 - z^2 & 3yz \\ 3xz & 3yz & 2z^2 - x^2 - y^2 \end{pmatrix} dV . \tag{21.57}$$

Note the resemblance between (21.55) and (21.52). For the dipole field the directional dependence is described by the single contraction $\left(\hat{\mathbf{r}} \cdot \mathbf{r}_g\right)$ whereas for the quadrupole field directional dependence is now given by the double contraction $(\hat{\mathbf{r}}\hat{\mathbf{r}} : \mathbf{T})$. This double contraction leads to a greater angular dependence of the quadrupole term than for the monopole term and the dipole term.

To find the angular dependence, we use that the inertia tensor $\mathbf{T}$ is a real symmetric 3×3 matrix. This matrix therefore has three orthogonal eigenvectors $\hat{\mathbf{v}}^{(i)}$ with corresponding eigenvalues λ_i. Using expression (13.45) this implies that the quadrupole moment tensor can be written as:

$$\mathbf{T} = \sum_{i=1}^{3} \lambda_i \hat{\mathbf{v}}^{(i)} \hat{\mathbf{v}}^{(i)}. \tag{21.58}$$

Problem i Use this result to show that the quadrupole term can be written as

$$V_{qua}(\mathbf{r}) = -\frac{G}{2r^3} \sum_{i=1}^{3} \lambda_i \cos^2 \Psi_i , \tag{21.59}$$

where the Ψ_i denote the angles between the eigenvectors $\hat{\mathbf{v}}^{(i)}$ and the observation direction $\hat{\mathbf{r}}$; see Figure 21.7 for the definition of these angles.

The directional dependence of (21.59) varies as $\cos^2 \Psi_i = (\cos 2\Psi_i + 1)/2$; this implies that the quadrupole field varies through two periods when Ψ_i increases from 0 to 2π. This contrasts with the monopole field, which does not depend on the direction at all, as well as with the dipole field that varies according to problem e as $\cos\theta$. There is actually a close connection between the different terms in the multipole expansion and spherical harmonics. This can be seen by comparing the multipole terms (21.47)–(21.49) with (21.36) for the gravitational potential. In (21.36), the different terms decay with distance as $r^{-(l+1)}$ and have an angular dependence $Y_{lm}(\theta, \varphi)$. Similarly, the multipole terms decay as r^{-1}, r^{-2},

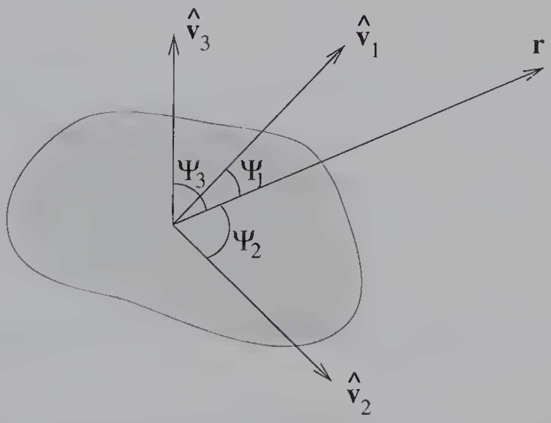

Fig. 21.7 Definition of the angles Ψ_i.

and r^{-3}, respectively, and depend on the direction as $\cos 0$, $\cos\theta$, and $\cos 2\Psi$ respectively.

21.7 Quadrupole field of the Earth

Let us now investigate what the multipole expansion implies for the gravity field of the Earth. The monopole term is by far the dominant term. It explains why an apple falls from a tree, why the Moon orbits the Earth and most other manifestations of gravity that we observe in daily life. The dipole term has in this context no physical meaning whatsoever. This can be seen from (21.52) which states that the dipole term only depends on the distance from the Earth's center of gravity to the origin of the coordinate system. Since we are free in choosing the origin, the dipole term can be made to vanish by choosing the origin of the coordinate system as the Earth's center of gravity. It is through the quadrupole field that some of the subtleties of the Earth's gravity field become manifest.

Problem a The quadrupole field vanishes when the mass distribution in the Earth is spherically symmetric. Show this by computing the inertia tensor **T** when $\rho = \rho(r)$.

The dominant departure of the shape of the Earth from spherical is the flattening of the Earth due to its rotation. If that is the case, then by symmetry one eigenvector of **T** must be aligned with the Earth's axis of rotation, and the two other eigenvectors must be perpendicular to the axis of rotation. By symmetry these other eigenvectors must correspond to equal eigenvalues. When we choose a coordinate system with the z-axis along the Earth's axis of rotation the eigenvectors are therefore given by the unit vectors $\hat{\mathbf{z}}$, $\hat{\mathbf{x}}$, and $\hat{\mathbf{y}}$ with eigenvalues λ_z, λ_x, and λ_y, respectively. The last two eigenvalues are identical because of the rotational symmetry around the Earth's axis of rotation: hence $\lambda_y = \lambda_x$.

Let us first determine the eigenvalues. Once the eigenvalues are known, the quadrupole moment tensor follows from (21.59). The eigenvalues could be found in the standard way by solving $\det(\mathbf{T}-\lambda\mathbf{I}) = 0$, but this is unnecessarily difficult. Once we know the eigenvectors the eigenvalues can easily be found from (21.58).

Problem b Take twice the inner product of (21.58) with the eigenvector $\hat{\mathbf{v}}^{(j)}$ to show that

$$\lambda_j = \hat{\mathbf{v}}^{(j)} \cdot \mathbf{T} \cdot \hat{\mathbf{v}}^{(j)}. \tag{21.60}$$

Problem c Use this with (21.57) to show that the eigenvalues are given by:

$$\left.\begin{aligned} \lambda_x &= \int \rho(\mathbf{r})\left(2x^2 - y^2 - z^2\right) dV, \\ \lambda_y &= \int \rho(\mathbf{r})\left(2y^2 - x^2 - z^2\right) dV, \\ \lambda_z &= \int \rho(\mathbf{r})\left(2z^2 - x^2 - y^2\right) dV. \end{aligned}\right\} \tag{21.61}$$

It is useful to relate these eigenvalues to the Earth's *moments of inertia*. The moment of inertia of the Earth around the z-axis is defined as:

$$C \equiv \int \rho(\mathbf{r})\left(x^2 + y^2\right) dV, \tag{21.62}$$

whereas the moment of inertia around the x-axis is defined as

$$A \equiv \int \rho(\mathbf{r})\left(y^2 + z^2\right) dV. \tag{21.63}$$

By symmetry the moment of inertia around the y-axis is given by the same moment A. These moments describe the rotational inertia around the coordinate axes as shown in Figure 21.8. The eigenvalues in (21.61) can be related to these moments of inertia. Because of the assumed axisymmetric density distribution, the integral of y^2 in (21.61) is equal to the integral of x^2. The eigenvalue λ_x in (21.61) is therefore given by: $\lambda_x = \int \rho(\mathbf{r})\left(x^2 - z^2\right) dV = \int \rho(\mathbf{r})\left(x^2 + y^2 - y^2 - z^2\right) dV = C - A$.

Problem d Apply a similar treatment to the other eigenvalues to show that:

$$\lambda_x = \lambda_y = C - A\,, \qquad \lambda_z = -2(C - A). \tag{21.64}$$

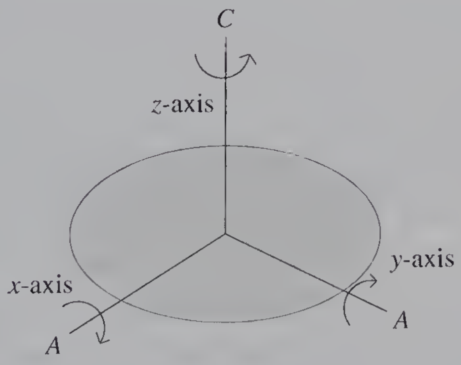

Fig. 21.8 Definition of the moments of inertia A and C for an Earth with cylindrical symmetry around the rotation axis.

Problem e Use these eigenvalues in (21.58) together with (21.55) and expression (4.7) for the unit vector $\hat{\mathbf{r}}$ to show that the quadrupole field is given by:

$$V_{qua}(\mathbf{r}) = \frac{G}{2r^3}(C - A)\left(3\cos^2\theta - 1\right). \tag{21.65}$$

The Legendre polynomial of order 2 is given by: $P_2^0(x) = \frac{1}{2}\left(3x^2 - 1\right)$. The quadrupole field can therefore be written as:

$$V_{qua}(\mathbf{r}) = \frac{G}{r^3}(C - A)\, P_2^0(\cos\theta). \tag{21.66}$$

The term $(C - A)$ denotes the difference in the moments of inertia of the Earth around the rotation axis and around an axis through the equator, see Figure 21.8. If the Earth were a perfect sphere, these moments of inertia would be identical and the quadrupole field would vanish. However, the rotation of the Earth causes the Earth to bulge at the equator. This departure from spherical symmetry is responsible for the Earth's quadrupole field.

If the Earth were spherical, the motion of satellites orbiting the Earth would satisfy Kepler's laws. The quadrupole term in the potential effects a measurable deviation of the trajectories of satellites from the orbits predicted by Kepler's laws. For example, if the potential is spherically symmetric, a satellite orbits in a fixed plane. The quadrupole field causes the plane in which the satellite orbits to precess slightly. Observations of the orbits of satellites can therefore be used to deduce the departure of the Earth's shape from spherical symmetry [59]. Using these techniques, it has been found that the difference $(C - A)$ in the moments of inertia has the numerical value [103]:

$$J_2 = \frac{(C - A)}{Ma^2} = 1.082626 \times 10^{-3}. \tag{21.67}$$

In this expression a is the radius of the Earth, and the term Ma^2 is a measure of the average moment of inertia of the Earth. Expression (21.67) states therefore that the relative departure of the mass distribution of the Earth from spherical symmetry is of the order 10^{-3}. This effect is small, but this number carries important information about the dynamics of our planet. In fact, the time derivative $\dot{J}_2$ of this quantity has been measured [124] as well! This quantity is of importance because the rotation rate of the Earth slowly decreases due to the braking effect of the tidal forces. Both the oceans and the solid Earth are subject to the tidal force that operates in the Earth–Moon system. There is a time lag in the response of both the ocean and the solid Earth and the tidal force that generates the tides. This time lag leads to a deceleration of the rotation of the Earth [103]; the Earth adjusts its shape to this deceleration. The measurement of $\dot{J}_2$ therefore provides important information about the response of the Earth to a time-dependent loading.

21.8 Fifth force

Gravity is the force in nature that was first understood by mankind through the discovery by Newton of the law of gravitational attraction. The reason that the gravitational force was understood first is that this force manifests itself in the macroscopic world in the motion of the Sun, Moon, and planets. Later the electromagnetic force, and the strong and weak interactions were discovered. This means that presently four forces are thought to be operative in nature. Of these four forces, the electromagnetic force and the weak nuclear force can now be described by a single unified theory.

In the 1980s, geophysical measurements of gravity suggested that the gravitational field behaves in a different way over geophysical length scales (between meters and kilometers) than over astronomical length scales ($>10^4$ km). This has led to the speculation that this discrepancy is due to a fifth force in nature. This speculation and the observations that fuelled this idea are clearly described by Fishbach and Talmadge [41]. The central idea is that in Newton's theory of gravity the gravitational potential generated by a point mass M is given by (21.38):

$$V_N(\mathbf{r}) = -\frac{GM}{r}. \tag{21.68}$$

The hypothesis of the fifth force presumes that a new potential should be added to this Newtonian potential: this fifth potential is given by

$$V_5(\mathbf{r}) = -\alpha\frac{GM}{r}e^{-r/\lambda}. \tag{21.69}$$

Note that this potential has almost the same form as the Newtonian potential $V_N(\mathbf{r})$, the main differences are that the fifth force decays exponentially with distance over a length λ and that it is weaker than Newtonian gravity by a factor α. This idea was prompted by measurements of gravity in mines, in the ice-cap of Greenland, on a 600 m high telecommunication tower, and by a number of other experimental results that seemed to disagree with the gravitational force that follows from the Newtonian potential $V_N(\mathbf{r})$.

Problem a Effectively, the fifth force leads to a change of the gravitational constant G with distance. Compute the gravitational acceleration $\mathbf{g}(r)$ for the combined potential $V_N + V_5$ by taking the gradient and write the result as

$-G(r)M\hat{\mathbf{r}}/r^2$ to show that the effective gravitational constant is given by:

$$G(r) = G\left[1 + \alpha\left(1 + \frac{r}{\lambda}\right)e^{-r/\lambda}\right]. \tag{21.70}$$

The fifth force thus effectively leads to a change of the gravitational constant over a characteristic distance λ. This effect is small: in 1991 the value of α was estimated to be less than 10^{-3} for all estimates of λ longer than 1 cm [41].

In doing geophysical measurements of gravity, one has to correct for perturbing effects such as the topography of the Earth's surface and density variations within the Earth's crust. It has now been shown that the uncertainties in these corrections are much larger than the observed discrepancy between the gravity measurements and Newtonian gravity [80]. This means that the issue of the fifth force seems be closed for the moment, and that the physical world appears again to be governed by four fundamental forces.

22

Cartesian tensors

In physics and mathematics, coordinate transformations play an important role because many problems are much simpler when a suitable coordinate system is used. Furthermore, the requirement that physical laws do not change under certain transformations imposes constraints on the physical laws. An example of this is presented in Section 22.11 where it is shown that the fact that the pressure in a fluid is isotropic follows from the requirement that some physical laws may not change under a rotation of the coordinate system. In this chapter it is shown how the change of vectors and matrices under coordinate transformations is derived. The derived transformation properties can be generalized to other mathematical objects which are called tensors. In this chapter, only transformations of rectangular coordinate systems are considered. Since these coordinate systems are called Cartesian coordinate systems, the associated tensors are called Cartesian tensors. The transformation properties of tensors in Cartesian and curvilinear coordinate systems are described in detail by Butkov [24] and Riley *et al.* [87].

22.1 Coordinate transforms

In this section we consider the transformation of a coordinate system in two dimensions. In Figure 22.1 an old coordinate system with coordinates x^{old} and y^{old} is shown. The unit vectors along the old coordinate axis are denoted by $\hat{\mathbf{e}}^{x,old}$ and $\hat{\mathbf{e}}^{y,old}$. In a coordinate transformation, these old unit vectors are transformed to new unit vectors $\hat{\mathbf{e}}^{x,new}$ and $\hat{\mathbf{e}}^{y,new}$, respectively. The matrix that maps the old unit vectors onto the new unit vectors is denoted by $\mathbf{C}^{-1}$, hence

$$\mathbf{C}^{-1}\hat{\mathbf{e}}^{x,old} = \hat{\mathbf{e}}^{x,new}, \qquad \mathbf{C}^{-1}\hat{\mathbf{e}}^{y,old} = \hat{\mathbf{e}}^{y,new}. \tag{22.1}$$

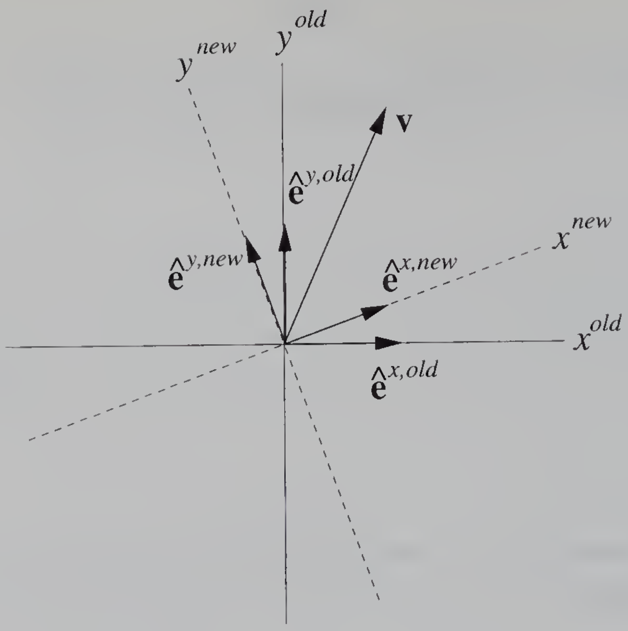
y^{old}
y^{new}
$\mathbf{v}$
$\hat{\mathbf{e}}^{y,old}$
x^{new}
$\hat{\mathbf{e}}^{y,new}$
$\hat{\mathbf{e}}^{x,new}$
x^{old}
$\hat{\mathbf{e}}^{x,old}$

Fig. 22.1 Definition of vectors in the original and transformed coordinate systems.

The fact that we use the inverse of the matrix has no significance; the only reason we do so is that in the following sections we will mostly use the inverse of $\mathbf{C}^{-1}$ which is equal to $\mathbf{C}$. The elements of the matrix $\mathbf{C}^{-1}$ are labeled as follows

$$\mathbf{C}^{-1} = \begin{pmatrix} C_{11}^{-1} & C_{12}^{-1} \\ C_{21}^{-1} & C_{22}^{-1} \end{pmatrix}, \tag{22.2}$$

where it is understood that C_{ij}^{-1} is not the inverse of the element C_{ij}, but is the i, j-element of the matrix $\mathbf{C}^{-1}$.

There is a close relation between the new basis vectors $\hat{\mathbf{e}}^{x,new}$ and $\hat{\mathbf{e}}^{y,new}$ and the matrix $\mathbf{C}^{-1}$. To see this we first use that in the old coordinate system the unit vector $\hat{\mathbf{e}}^{x,old}$ is the unit vector in the x-direction and that $\hat{\mathbf{e}}^{y,old}$ is the unit vector in the y-direction:

$$\hat{\mathbf{e}}^{x,old} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \qquad \hat{\mathbf{e}}^{y,old} = \begin{pmatrix} 0 \\ 1 \end{pmatrix}. \tag{22.3}$$

Problem a Let the matrix $\mathbf{C}^{-1}$ act on these expressions and use expression (22.1) to show that the first column of $\mathbf{C}^{-1}$ is given by $\hat{\mathbf{e}}^{x,new}$ while the second column is equal to $\hat{\mathbf{e}}^{y,new}$:

$$\hat{\mathbf{e}}^{x,new} = \begin{pmatrix} C_{11}^{-1} \\ C_{21}^{-1} \end{pmatrix}, \qquad \hat{\mathbf{e}}^{y,new} = \begin{pmatrix} C_{12}^{-1} \\ C_{22}^{-1} \end{pmatrix}. \tag{22.4}$$

Problem b Use (22.2) to show that this implies that the matrix $\mathbf{C}^{-1}$ can also be written as

$$\mathbf{C}^{-1} = \left(\hat{\mathbf{e}}^{x,new} \; \hat{\mathbf{e}}^{y,new} \right) . \tag{22.5}$$

In other words, the transformed basis vectors are the columns of the matrix that describes the coordinate transform.

Let us now consider a vector $\mathbf{v}$ as shown in Figure 22.1. This vector is a physical quantity that denotes, for example, the position of an object relative to some specified origin. This vector is independent of any coordinate system since it is a physical quantity. However, when we describe this vector we use a system of basis vectors to give the components of this vector in directions defined by the basis vectors. For example, using the old coordinate vectors $\hat{\mathbf{e}}^{x,old}$ and $\hat{\mathbf{e}}^{y,old}$ the vector $\mathbf{v}$ can be decomposed into its components v_x^{old} and v_y^{old}:

$$\mathbf{v} = v_x^{old} \hat{\mathbf{e}}^{x,old} + v_y^{old} \hat{\mathbf{e}}^{y,old} . \tag{22.6}$$

This same vector can also be decomposed into its components along the new coordinate system

$$\mathbf{v} = v_x^{new} \hat{\mathbf{e}}^{x,new} + v_y^{new} \hat{\mathbf{e}}^{y,new} . \tag{22.7}$$

The vector $\mathbf{v}$ is the same in each of these expressions. It is the goal of this section to derive the relation between the old components v^{old} and the new components v^{new}.

Problem c Use that (22.6) and (22.7) are expressions for the same vector and insert (22.3) and (22.4) for the unit vectors to show that the old and the new components are related by

$$v_x^{old} \begin{pmatrix} 1 \\ 0 \end{pmatrix} + v_y^{old} \begin{pmatrix} 0 \\ 1 \end{pmatrix} = v_x^{new} \begin{pmatrix} C_{11}^{-1} \\ C_{21}^{-1} \end{pmatrix} + v_y^{new} \begin{pmatrix} C_{12}^{-1} \\ C_{22}^{-1} \end{pmatrix} . \tag{22.8}$$

Problem d Rewrite this result to derive that

$$\left. \begin{aligned} C_{11}^{-1} v_x^{new} + C_{12}^{-1} v_y^{new} &= v_x^{old} , \\ C_{21}^{-1} v_x^{new} + C_{22}^{-1} v_y^{new} &= v_y^{old} . \end{aligned} \right\} \tag{22.9}$$

At this point it is useful to introduce the vectors $\mathbf{v}^{old}$ and $\mathbf{v}^{new}$ with the components of the vector in the old and new coordinate system respectively:

$$\mathbf{v}^{old} \equiv \begin{pmatrix} v_x^{old} \\ v_y^{old} \end{pmatrix} , \qquad \mathbf{v}^{new} \equiv \begin{pmatrix} v_x^{new} \\ v_y^{new} \end{pmatrix} . \tag{22.10}$$

This harmless-looking definition belies a subtle complication. We now have three notations for the same vector. The notation $\mathbf{v}$ denotes the physical vector that is defined independently of any coordinate system, while the vectors $\mathbf{v}^{old}$ and $\mathbf{v}^{new}$ give the *representation* of this vector with respect to the old and the new coordinate systems, respectively. It is crucial to separate these three different vectors.

Problem e Use the rules of matrix–vector multiplication to show that (22.9) implies that

$$\mathbf{C}^{-1}\mathbf{v}^{new} = \mathbf{v}^{old}, \tag{22.11}$$

and derive from this that

$$\mathbf{v}^{new} = \mathbf{C}\mathbf{v}^{old}. \tag{22.12}$$

This is the desired result because this expression prescribes how the old vector $\mathbf{v}^{old}$ is mapped in the coordinate transform onto the new vector $\mathbf{v}^{new}$. It is interesting to compare this expression with (22.1) which states that the old unit vectors are mapped onto the new basis vectors by the *inverse* matrix $\mathbf{C}^{-1}$. There is a good reason why the basis vectors transform with the inverse of the transformation matrix of the components. According to (22.6) the physical vector $\mathbf{v}$ is the product of the components of the vector and the basis vectors: $\mathbf{v} = v_x\hat{\mathbf{e}}^x + v_y\hat{\mathbf{e}}^y$. Since the physical vector $\mathbf{v}$ by definition is not changed by the coordinate transform, the *product* $v_x\hat{\mathbf{e}}^x + v_y\hat{\mathbf{e}}^y$ of the components and the basis vectors is invariant. This can only be the case when the transformation rule for the components is the inverse of the transformation rule of the basis vectors.

The treatment in this section was for a two-dimensional coordinate transform. Each step of the argument can be generalized to coordinate transforms in more than two dimensions.

22.2 Unitary matrices

In the previous section, it was tacitly assumed that the matrix $\mathbf{C}$ was defined in such a way that the new basis vectors $\hat{\mathbf{e}}^{x,new}$ and $\hat{\mathbf{e}}^{y,new}$ were of unit length. For a general coordinate transform $\mathbf{C}$ this is not true. We are, however, interested in orthonormal coordinate systems only: these are coordinate systems in which the basis vectors have unit length and are mutually orthogonal.

Problem a Let the unit vectors in a coordinate system be denoted by $\hat{\mathbf{e}}^i$. Show that for an orthonormal coordinate system

$$\left(\hat{\mathbf{e}}^i \cdot \hat{\mathbf{e}}^j\right) = \delta_{ij}, \tag{22.13}$$

where δ_{ij} is the Kronecker delta which was defined in (11.15) as

$$\delta_{ij} = \begin{cases} 1 & \text{when} \quad i = j \\ 0 & \text{when} \quad i \neq j \end{cases} \tag{11.15}$$

When the old coordinate system and the new coordinate system are orthonormal, both coordinate systems must satisfy (22.13). The condition is certainly satisfied when the inner product of two vectors does not change under the coordinate transformation. Such coordinate transformations are called *unitary transformations*. To be more specific, let **u** and **v** be two vectors in the old coordinate system that are mapped by the coordinate transform **C** to new vectors $\mathbf{u}' = \mathbf{C}\mathbf{u}$ and $\mathbf{v}' = \mathbf{C}\mathbf{v}$, respectively. The coordinate transform **C** is called unitary when $(\mathbf{u}' \cdot \mathbf{v}') = (\mathbf{u} \cdot \mathbf{v})$, or equivalently

$$(\mathbf{C}\mathbf{u} \cdot \mathbf{C}\mathbf{v}) = (\mathbf{u} \cdot \mathbf{v}). \tag{22.14}$$

Problem b Show that the fact that the length of a vector and the angle between vectors is preserved implies that the inner product of two vectors is preserved as well. Examples of unitary coordinate transformations are rotations, reflections in a plane, or combinations of these. Draw a number of figures to convince yourself that these coordinate transformations do not change the length of the vector and the angle between two vectors so that these transformations are indeed unitary.

Problem c The requirement (22.14) can be used to impose a constraint on **C**. Write the matrix products and the inner product in (22.14) in terms of the components of the matrix **C** and the vectors **u** and **v** to derive that it is equivalent to

$$\sum_{i,j,k} C_{ij} u_j C_{ik} v_k = \sum_i u_i v_i . \tag{22.15}$$

Writing out a matrix expression in its components has an important advantage. In general one cannot interchange the order of two matrices, because in general the matrix product **AB** differs from the matrix product **BA**.

Problem d Verify this statement by taking

$$\mathbf{A} = \begin{pmatrix} 0 & 1 & 0 \\ -1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix} \quad \text{and} \quad \mathbf{B} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & -1 & 0 \end{pmatrix} \tag{22.16}$$

and showing that $\mathbf{AB} \neq \mathbf{BA}$.

Problem e Geometrically the matrix **A** represents a rotation around the z-axis through 90 degrees while **B** represents a rotation around the x-axis through 90 degrees. Take this book and carry out these rotations in the two different orders to see that the final orientation of the book is indeed different in each case.

This means that in general one cannot interchange the order of matrices in a product. However, when a term $A_{ij}B_{jk}$ occurs in a summation, the terms A_{ij} and B_{jk} refer to individual matrix elements. These matrix elements are regular numbers that can be interchanged: $A_{ij}B_{jk} = B_{jk}A_{ij}$. This means that the decomposition of a matrix equation into its components allows us to move the individual components around in any way that we like.

In the following we will use the transpose $\mathbf{C}^T$ of a matrix **C** extensively. This is the matrix in which the rows and the columns are interchanged, so that

$$C_{ij}^T \equiv C_{ji}. \tag{22.17}$$

Problem f Use the results in this section to show that (22.15) can be written as

$$\sum_{i,j,k} C_{ji}^T C_{ik} u_j v_k = \sum_i u_i v_i. \tag{22.18}$$

At this point it is important to realize that it is immaterial how we label the summation index. The right-hand side of (22.18) can just as well be written as $\sum_j u_j v_j$ because this sum and $\sum_i u_i v_i$ both mean: "take the corresponding elements of **u** and **v**, multiply them and sum over all components."

Problem g Show that (22.18) is equivalent to

$$\sum_{j,k} \left(\sum_i C_{ji}^T C_{ik} \right) u_j v_k = \sum_j u_j v_j. \tag{22.19}$$

Problem h Use the definition (11.15) of the Kronecker delta to show that $\sum_j u_j v_j = \sum_{j,k} \delta_{jk} u_j v_k$.

Problem i Use this result to show that **C** is unitary when it satisfies the following condition

$$\left(\sum_i C_{ji}^T C_{ik} \right) = \delta_{jk}. \tag{22.20}$$

At this point we use that δ_{jk} is just the j,k element of the identity matrix $\mathbf{I}$ because the elements of this matrix are equal to 1 on the diagonal ($j = k$) and equal to zero off the diagonal ($j \neq k$). In matrix notation, (22.20) for a unitary matrix therefore implies that

$$\mathbf{C}^T\mathbf{C} = \mathbf{I}. \tag{22.21}$$

Problem j Show that this implies that the inverse of a unitary matrix is equal to its transpose:

$$\mathbf{C}^{-1} = \mathbf{C}^T. \tag{22.22}$$

This result can save you a lot of work. Suppose you need to invert a matrix, and suppose that you suspect that the matrix is unitary, then according to (22.22) the inverse matrix is equal to the transpose. You can verify whether this is indeed the case by multiplying this "guess" by the original matrix $\mathbf{C}$ to check whether the result is equal to the identity matrix; this amounts to applying (22.21).

Problem k The matrix corresponding to a rotation over an angle φ is given by

$$\mathbf{R} = \begin{pmatrix} \cos\varphi & \sin\varphi \\ -\sin\varphi & \cos\varphi \end{pmatrix}. \tag{22.23}$$

Compute the inverse of this matrix with a minimal amount of work. (This is also called the principle of maximal laziness.)

22.3 Shear or dilatation?

As an example of the importance of coordinate transforms we consider a geologist who observes a rock sample that is subjected to a shear as shown in Figure 22.2. The height and the width of the rock are of unit length. In the shear motion the upper side of the rock is displaced rightward over a distance d while the lower side of the rock is displaced leftward over a distance $-d$.

Problem a Use Figure 22.2 to show that the original unit vectors change during the deformation as

$$\begin{pmatrix} 1 \\ 0 \end{pmatrix} \rightarrow \begin{pmatrix} 1 \\ 0 \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} 0 \\ 1 \end{pmatrix} \rightarrow \begin{pmatrix} d \\ 1 \end{pmatrix}. \tag{22.24}$$

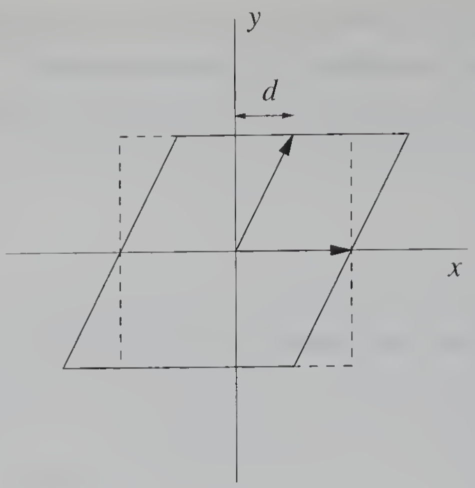

Fig. 22.2 The shear of a rock as seen by the geologist before (dashed lines) and after (solid lines) the shear.

Problem b Let the deformation be described by a 2 × 2 matrix **D**. Show that this matrix is given by

$$\mathbf{D} = \begin{pmatrix} 1 & d \\ 0 & 1 \end{pmatrix}. \tag{22.25}$$

Hint: check that this produces the deformation described by (22.24).

The geologist is in general interested in the internal deformation of the rock. However, the transformation shown in Figure 22.2 entails both a deformation within the rock sample as well as a rotation of the sample. We first need to unravel these two contributions. In the following we only considered very small shear ($d \ll 1$), the associated rotation is therefore also small.

Problem c Use a Taylor expansion of (22.23) to show that the rotation matrix over an infinitesimal angle φ is given by

$$\mathbf{R} = \mathbf{I} + \begin{pmatrix} 0 & \varphi \\ -\varphi & 0 \end{pmatrix}. \tag{22.26}$$

The shear described by (22.25) can be decomposed into this rotation plus an internal deformation. To make this explicit we decompose the matrix **D** into the matrix on the right-hand side of (22.26) which accounts for the rotation and a new matrix **E**:

$$\mathbf{D} = \begin{pmatrix} 0 & \varphi \\ -\varphi & 0 \end{pmatrix} + \mathbf{E}. \tag{22.27}$$

At this point we need to find the parameter φ that characterizes the rotational component. It follows from (22.26) that the difference between the 12-component and the 21-component of the matrix is equal to 2φ. According to (22.25) this difference is given by $D_{12} - D_{21} = d$; since this is equal to 2φ the rotation angle is given by $\varphi = d/2$.

Problem d Use the expressions (22.25), (22.27), and the relation $\varphi = d/2$ to show that the deformation component of the transformation in Figure 22.2 is given by

$$\mathbf{E} = \begin{pmatrix} 1 & d/2 \\ d/2 & 1 \end{pmatrix}. \tag{22.28}$$

Note that this matrix is symmetric $(E_{ij} = E_{ji})$, whereas the matrix $\mathbf{R}$ in (22.26) is antisymmetric $(R_{ij} = -R_{ji})$. The matrix $\mathbf{D}$ has thus been decomposed into a symmetric part and an antisymmetric part. This corresponds to a decomposition of the deformation into a rotational component plus a deformation. We have seen this before in Chapter 7 where it was shown that the curl of a vector field in general has two contributions: rotation and shear. We return to this issue in Section 22.9.

Problem e We now focus on the internal deformation of the sample. Show that Figure 22.3 displays the deformation $\mathbf{E}$ of the rock. Hint: determine first what happens to the unit vectors in the x- and y-directions when they are multiplied with the matrix $\mathbf{E}$.

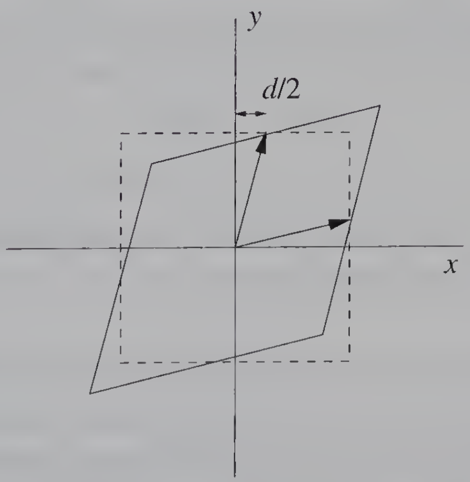

Fig. 22.3 The same deformation as in the previous figure when the rotational component is removed. The undeformed sample is shown by the dashed lines, the deformed sample by the solid lines.

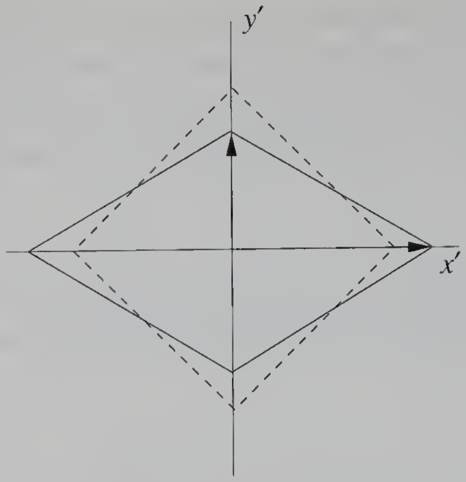

Fig. 22.4 The deformation of the previous figure seen by a geologist who uses a coordinate system that is rotated through 45 degrees.

Let us now consider a second geologist who is studying the same deforming rock but who is using a coordinate system that is rotated through 45 degrees with respect to the coordinate system used by the geologist who views the world as seen in Figure 22.2. We ignore the effect of the rotation altogether because it does not lead to internal deformation of the rock. The x'-axis of the second geologist is aligned along the line $x = y$, whereas the y'-axis is aligned along the line $x = -y$. Hence the basis vectors of the second geologist are given by

$$\hat{\mathbf{e}}^{x'} = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \qquad \hat{\mathbf{e}}^{y'} = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ -1 \end{pmatrix}. \tag{22.29}$$

Problem f The simplest way to determine how the second geologist views the deformation is to determine what happens when **E** acts on the basis vectors of the second geologist. Show that this gives

$$\mathbf{E}\hat{\mathbf{e}}^{x'} = (1 + d/2)\,\hat{\mathbf{e}}^{x'}, \qquad \mathbf{E}\hat{\mathbf{e}}^{y'} = (1 - d/2)\,\hat{\mathbf{e}}^{y'}. \tag{22.30}$$

Problem g Use this result to show that the second geologist sees the deformation process as depicted in Figure 22.4.

Note that the unit vectors of the second geologist are the eigenvectors of the matrix **E**.

The second geologist would describe the deformation process seen in Figure 22.4 as follows. The sample is extended in the x'-direction because the unit vector $\hat{\mathbf{e}}^{x'}$ is mapped onto itself with an amplification factor $(1 + d/2)$ that is greater than 1; in the y'-direction the sample is compressed because the unit vector $\hat{\mathbf{e}}^{y'}$ is mapped onto itself with an amplification factor $(1 - d/2)$ that is smaller than 1. This means that (apart from a rotation) the first geologist describes the transformation as pure shear whereas the second geologist describes the transformation as a combination of extension and compression. These two points of view must be equivalent. However, they can be reconciled only when we have techniques for switching from one coordinate system to another. This is the goal of tensor calculus. Before we derive the transformation rules for matrices and more general tensors, we first introduce a notational convention that turns out to be very convenient.

22.4 Summation convention

In this section the summation convention is introduced; this convention is sometimes also referred to as the "Einstein summation convention." Let us consider the multiplication of the matrix **A** by a vector **x** that gives a new vector **y** = **Ax**.

Problem a This is a vectorial notation; show that in component form this expression should be written as

$$\sum_j A_{ij} x_j = y_i . \tag{22.31}$$

You have seen in Section 22.2 that in matrix and vector calculations one has to carry out a lot of summations. It is also shown in that section that working with the expressions in component form has advantages over working with the more abstract vector/matrix notation. We could save a lot of work by leaving out the summation signs in the equations. This is exactly what the summation convention does:

> *In the summation convention one leaves out the summation signs ($\sum$) with the understanding that one sums over any repeated index on either side of the equation.*

What does this mean for expression (22.31)? In this expression we sum over the index j. Note that this index appears twice on the left-hand side. According to the summation convention we would write this expression as

$$A_{ij} x_j = y_i . \tag{22.32}$$

It takes some time to get used to this notation. It should be kept in mind that under the summation convention the identity (22.32) does *not* imply that this expression holds for a single value of j. For example, (22.32) does not mean that $A_{i2}x_2 = y_i$ because a summation over the index j is implied rather than that j can take any fixed value $j = 1$ or $j = 2$ or $j = 3$.

As another example of the summation convention consider the inner product of two vectors. Normally one would write this as $(\mathbf{u}\cdot\mathbf{v}) = \sum_i u_i v_i$; using the summation convention one writes it as $(\mathbf{u}\cdot\mathbf{v}) = u_i v_i$. Note that the name of the summation index is irrelevant. This expression could therefore also be written as $(\mathbf{u}\cdot\mathbf{v}) = u_j v_j$. The matrix equation $\mathbf{C}^T\mathbf{C} = \mathbf{I}$ is written in component form as $\sum_k C^T_{ik} C_{kj} = \sum_k C_{ik} C_{kj} = \delta_{ij}$; in the summation convention this is written as $C^T_{ik} C_{kj} = C_{ik} C_{kj} = \delta_{ij}$.

Problem b Write out the matrix product $\mathbf{AI}$ (with $\mathbf{I}$ as the identity matrix) in component form, and show that the result can be written as: $(\mathbf{AI})_{ij} = A_{ik}\delta_{kj}$.

In the last term a summation over the index k is implied. The Kronecker delta is always zero, except when $k = j$. This means that the only nonzero contribution from the k-summation comes from the term $k = j$, so that $(\mathbf{AI})_{ij} = A_{ik}\delta_{kj} = A_{ij}$. This is simply the identity $\mathbf{AI} = \mathbf{A}$ in component form.

Problem c Show that $v^2 = v_i v_i$ where v is the length of the vector $\mathbf{v}$.

Problem d The Laplacian is the divergence of the gradient, use this to write the Laplacian in the following form

$$\nabla^2 f = \partial_i \partial_i f, \tag{22.33}$$

with $\partial_i \equiv \partial/\partial x_i$.

Problem e Use the summation convention to rewrite expression (11.55) as

$$\frac{\partial(\rho v_i)}{\partial t} + \partial_j(\rho v_j v_i) = \mu \partial_k \partial_k v_i + F_i. \tag{22.34}$$

Problem f In some applications one carries out summations over several indices. As an example consider the double sum $\sum_{i,j} A_{ij}\delta_{ji}$. Carry out the summation and show that the result can be written as

$$A_{ij}\delta_{ji} = A_{ii}. \tag{22.35}$$

Note that on the left-hand side a summation over i and j is implied whereas on the right-hand side a sum over i is implied. The term on the right-hand side is the sum of the diagonal elements of the matrix $\mathbf{A}$. This quantity plays an important role in a variety of applications and is called the *trace* of $\mathbf{A}$:

$$\mathrm{tr}\,\mathbf{A} \equiv \sum_i A_{ii} = A_{ii}. \tag{22.36}$$

One caveat should be made with the summation convention. Suppose that one writes $M_{ii} = 0$. According to the summation convention this means that the trace of $\mathbf{M}$ is equal to zero. However, one could also mean to express that all the diagonal elements of $\mathbf{M}$ are equal to zero: $M_{11} = M_{22} = M_{33} = 0$. This is an example where an expression can be ambiguous. It is important to state explicitly whether one uses the summation convention. When one does, one can indicate that one deviates from this convention in an equation by explicitly stating that no summation is implied. For example, when one wants to express that all the diagonal elements of $\mathbf{M}$ are equal to zero one could write this in the summation convention as:

$$M_{ii} = 0 \qquad \text{(no summation over } i\text{)}. \tag{22.37}$$

Problem g Use the properties of unitary matrices as derived in Section 22.2 to show that for a unitary matrix $\mathbf{C}$:

$$C_{ij} C_{ik} = \delta_{kj}. \tag{22.38}$$

22.5 Matrices and coordinate transforms

In this section we determine the transformation properties of matrices under a coordinate transform $\mathbf{C}$ as introduced in Section 22.1. In the remainder of this section we then restrict ourselves to unitary coordinate transformations. Let a matrix $\mathbf{D}$ map a vector $\mathbf{x}$ to a vector $\mathbf{y}$

$$\mathbf{y} = \mathbf{D}\mathbf{x}. \tag{22.39}$$

Consider this same operation in a new coordinate system and let us use a prime to denote the corresponding quantities is the new coordinate system. According to (22.12) the vector $\mathbf{x}$ in the new coordinate system is given by $\mathbf{x}' = \mathbf{C}\mathbf{x}$, and a similar expression holds for $\mathbf{y}$. Expression (22.39) is given in the new coordinate system by

$$\mathbf{y}' = \mathbf{D}'\mathbf{x}'. \tag{22.40}$$

In this section we determine the relation between the matrix $\mathbf{D}'$ in the new coordinate system and the old matrix $\mathbf{D}$.

Problem a Use (22.12) to show that (22.40) can be written as

$$\mathbf{C}\mathbf{y} = \mathbf{D}'\mathbf{C}\mathbf{x}. \tag{22.41}$$

Problem b Multiply this expression on the left by $\mathbf{C}^{-1}$, compare the result with (22.39) to derive that $\mathbf{D} = \mathbf{C}^{-1}\mathbf{D}'\mathbf{C}$. Multiply this last expression on the left and the right by suitable matrices to derive that

$$\mathbf{D}' = \mathbf{C}\mathbf{D}\mathbf{C}^{-1}. \tag{22.42}$$

This is the general transformation rule for matrices under a coordinate transform. It looks different from the transformation rule (22.12) for vectors because in (22.42) the inverse $\mathbf{C}^{-1}$of the coordinate transformation appears as well. However, for unitary coordinate transforms one can eliminate this inverse and write (22.42) in a way that is similar to (22.12).

Problem c Use that $\mathbf{C}$ is unitary and rewrite (22.42) in component form as

$$D'_{ij} = \sum_{k,l} C_{ik} C_{jl} D_{kl}. \tag{22.43}$$

(Remember that you cannot interchange the order of matrices in a multiplication, but that you can change the order of matrix *elements* that are multiplied.)

Problem d Redo the calculation in the last problem using the summation convention at every step and rewrite the preceding result as

$$D'_{ij} = C_{ik} C_{jl} D_{kl}. \tag{22.44}$$

Problem e To compare this result with the transformation rule (22.12) for vectors write $\mathbf{v}^{old} = \mathbf{v}$, $\mathbf{v}^{new} = \mathbf{v}'$ and use the summation convention to rewrite (22.12) using the summation convention as

$$v'_i = C_{ij} v_j. \tag{22.45}$$

Applying the matrix transformation (22.44) may appear to be complex to you. The easiest way to apply it in practice is to write it as

$$\mathbf{D}' = \mathbf{C}\mathbf{D}\mathbf{C}^T. \tag{22.46}$$

This expression is identical to (22.42) with the only exception that it uses that $\mathbf{C}$ is unitary: $\mathbf{C}^{-1} = \mathbf{C}^T$.

Problem f According to (22.23) a rotation through 45 degrees of the coordinate system is described by the coordinate transform

$$\mathbf{C} = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 1 \\ -1 & 1 \end{pmatrix} . \tag{22.47}$$

Check that this coordinate transformation transforms the matrix **E** in (22.28) to the matrix

$$\mathbf{E}' = \begin{pmatrix} 1 + d/2 & 0 \\ 0 & 1 - d/2 \end{pmatrix} . \tag{22.48}$$

Show also that this matrix describes the deformation shown in Figure 22.4.

22.6 Definition of a tensor

Expressions (22.44) and (22.45) allow us to see a resemblance between the transformation rule (22.44) for a matrix and the rule (22.45) for a vector. In both equations the old vector or matrix is multiplied by the matrix **C**; for the vector this happens once and for the matrix it happens twice. The *first* index of the quantity on the left-hand side (i in both cases) is the first index of the *first* matrix **C** on the right-hand side as indicated by the arrows in the following expressions:

$$v'_{\underset{\Uparrow}{i}} = C_{\underset{\Uparrow}{i}j} v_j , \qquad D'_{\underset{\Uparrow}{i}j} = C_{\underset{\Uparrow}{i}k} C_{jl} D_{kl} .$$

The *second* index of $\mathbf{D}'$ in (22.44) corresponds to the first index of the *second* term **C** as indicated by the arrows:

$$D'_{i\underset{\Uparrow}{j}} = C_{ik} C_{\underset{\Uparrow}{j}l} D_{kl} .$$

In both (22.44) and (22.45) the *second* index of each term **C** also occurs in the term **D** or **v** on which this matrix acts as shown by the arrows in the following expressions:

$$v'_i = C_{i\underset{\Uparrow}{j}} v_{\underset{\Uparrow}{j}} , \qquad D'_{ij} = C_{i\underset{\Uparrow}{k}} C_{jl} D_{\underset{\Uparrow}{k}l} , \qquad D'_{ij} = C_{ik} C_{j\underset{\Uparrow}{l}} D_{k\underset{\Uparrow}{l}} .$$

Note that according to the summation convention, a summation over these repeated indices is implied.

Some quantities that we use are not labeled by any index. Examples are temperature and pressure. Such a quantity is a pure number and it is called a *scalar*. A vector has a direction and is labeled with one index and a matrix has two indices. There is no reason not to define objects which are characterized with an arbitrary number of indices. In general a physical quantity can have any number of subscripts, and a quantity with n subscripts can be denoted as $T_{i_1 i_2 \cdots i_n}$. We call such an object a tensor when it follows a transformation rule similar to (22.44) and (22.45) for a matrix and a vector, respectively.

Definition The quantity $T_{i_1 i_2 \cdots i_n}$ is called a tensor of rank n when it transforms under a unitary coordinate transform $\mathbf{C}$ with the following transformation rule:

$$T'_{i_1 i_2 \cdots i_n} = C_{i_1 j_1} C_{i_2 j_2} \cdots C_{i_n j_n} T_{j_1 j_2 \cdots j_n}. \tag{22.49}$$

Note that the summation convention is used so a summation over the indices $j_1, j_2, \ldots, j_n$ is implied.

Problem a Convince yourself that the transformation rules (22.44) and (22.45) for a matrix and a vector, respectively, are special cases of this definition for the values $n = 2$ and $n = 1$, respectively.

This means that a vector that transforms according to (22.45) is a tensor of rank one and a matrix that transforms according to (22.44) is a tensor of rank two. Note that a scalar does not change under coordinate transformations. It has zero subscripts and one can say that when one applies the coordinate transformation (22.49) the coordinate transform $\mathbf{C}$ is applied zero times, for this reason one calls a scalar a tensor of rank zero.

Let us consider some examples of tensors. The position vector $\mathbf{r}$ is a tensor of rank one, as follows from the transformation rule (22.12) that you derived in Section 22.1. From this it follows that the velocity vector $\mathbf{v}$ is a tensor of rank one provided that the coordinate transformation $\mathbf{C}$ does not depend on time.

Problem b Differentiate the transformation law $x'_i = C_{ij} x_j$ for the position vector with respect to time and use the result to show that the velocity indeed transforms as a tensor of rank one.

Problem c Show that acceleration is a tensor of rank one.

Now we can use Newton's law $\mathbf{F} = m\mathbf{a}$ and that the mass m is a scalar. Since the acceleration is a tensor of rank one, the force must also be a tensor of rank one.

It is interesting to see what happens when one works in a rotating coordinate system. In that case the transformation **C** which transforms the fixed system to the rotating system depends on time.

Problem d Go through the steps of problems b and c and show that in that case the acceleration transforms as

$$a_i' = C_{ij} a_j + 2\dot{C}_{ij} v_j + \ddot{C}_{ij} x_j, \tag{22.50}$$

where the dot denotes a time derivative.

This means that for such a time-dependent coordinate transform the acceleration does *not* transform as a vector of rank one. Note that additional terms appear that are proportional to the velocity and the position vector. These correspond to the Coriolis force and the centrifugal force as treated in Section 13.3.

Another example of a tensor of rank one is the gradient vector $\nabla_i = \partial/\partial x_i$. The chain rule of differentiation states that

$$\nabla_i' = \frac{\partial}{\partial x_i'} = \frac{\partial x_j}{\partial x_i'} \frac{\partial}{\partial x_j} = \frac{\partial x_j}{\partial x_i'} \nabla_j, \tag{22.51}$$

where to conform with the summation convention a summation over j is implied.

Problem e Use the transformation law of the position vector to derive that $\partial x_j/\partial x_i' = C_{ji}^{-1}$, and use the property that **C** is a unitary matrix to derive that

$$\nabla_i' = C_{ij} \nabla_j. \tag{22.52}$$

In other words, the gradient vector is a tensor of rank one.

As an example of a tensor of rank two we consider the identity matrix whose elements are the Kronecker delta δ_{ij}. At first sight you might believe that since the elements of the Kronecker delta are simply equal to zero or one, it is a scalar. However, the identity matrix really transforms like a tensor of rank two.

Problem f To see this, assume that the identity matrix follows the transformation rule (22.44) and use the property of the Kronecker delta to derive that

$$I_{ij}' = C_{ik} C_{jl} \delta_{kl} = C_{ik} C_{jk}. \tag{22.53}$$

Problem g Write the last term as C^T_{kj} and use the property that **C** is unitary to derive that

$$I'_{ij} = \delta_{ij}. \tag{22.54}$$

In other words, the identity matrix transforms as a tensor of rank two and yet it takes the same form in every coordinate system!

22.7 Not every vector is a tensor

At this point you may think that any object with n-components is a tensor of rank one. However, keep in mind that not every vector transforms according to the transformation rule (22.49) for a tensor. Let us first consider an obvious example. One might be interested in the distribution of shoe sizes of students. In such a study one would measure the shoe size of each student in the group and could form a vector of the data that one obtains:

$$\mathbf{d} \equiv \begin{pmatrix} \text{shoe size of Marie} \\ \text{shoe size of Peter} \\ \text{shoe size of Klaas} \\ \vdots \end{pmatrix}. \tag{22.55}$$

This vector is not a tensor. There are two reasons for this. The coordinate transformation matrix **C** is by definition a square matrix with a dimension equal to the dimension of the coordinate system (usually 3 and for some applications 4, see Section 22.12). The vector **d** can have any dimension and the action of **C** in this vector is not defined. The second reason why **d** is not a tensor of rank one is that the shoe size of students does not depend on the coordinate system. In other words, all the elements of **d** are scalars, hence **d** is certainly not a tensor of rank one.

Another example of a vector that is not a tensor is the stress–displacement tensor that is used in the description of elastic wave propagation in elastic media [3]:

$$\mathbf{w} = \begin{pmatrix} u_y \\ \sigma_{yz} \end{pmatrix}, \tag{22.56}$$

where u_y and σ_{yz} are suitably chosen components of the displacement and the stress. This vector is not a tensor of rank one. It has a different dimension than **C**, hence the action of **C** on this vector is not defined. Furthermore the components of **w** are of different dimension, hence they can never be mixed in the transformation rule (22.49).

Let us now consider a more subtle example. At this point you might think that the magnetic field vector is a tensor of rank one. Consider the circular current in

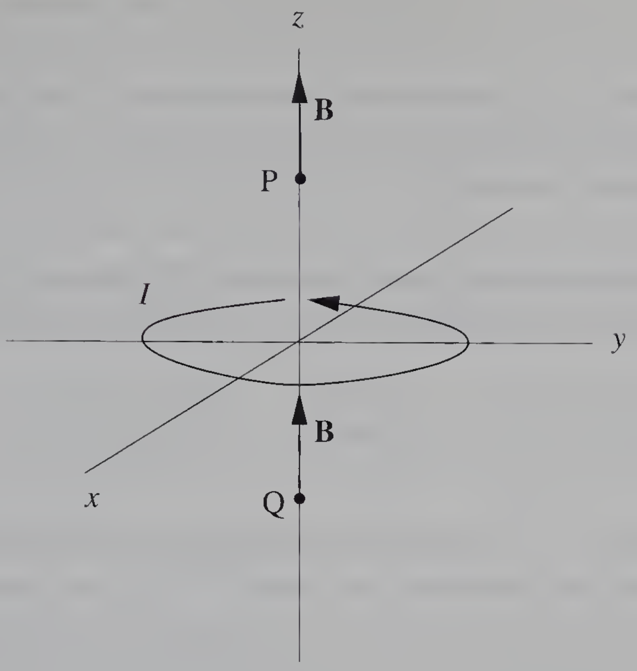

Fig. 22.5 A circular current that generates a magentic field **B**.

the (x, y)-plane shown in Figure 22.5 that generates a magnetic field. Given the direction of this current the magnetic field on the z-axis is oriented in the direction of the positive z-axis. This holds for the points P and Q in Figure 22.5.

Problem a Let us perform the coordinate transformation in which the system is reflected in the (x, y)-plane. Make a drawing of the transformed current and make sure you understand that the current has not changed in the coordinate transform.

Problem b Now that you know that the current has not changed, use Figure 22.5 to draw the associated magnetic field vector at the points P and Q.

Problem c Now add to your figure the image of the vectors **B** of Figure 22.5 after they have been subjected to a reflection in the (x, y)-plane.

If you made the drawing correctly, you will have found that the magnetic field that you obtained after transforming the original magnetic field vector had the opposite direction to the magnetic field that was generated by the current that had been subjected to the coordinate transform.

The only conclusion is that the magnetic field does not behave like a tensor of rank one under coordinate transforms: it does not transform according to the transformation rule (22.45). In fact, the magnetic field behaves in a different way. It is an example of a *pseudo-tensor* indicating that it shares some of the properties of the tensors that we have explored, but that its transformation rules differ from the transformation rule (22.45) that we have been used to. More details about pseudo-tensors can be found in reference [87]. The main thing to remember at this point is that not every vector is a tensor!

The next example shows that not every tensor is Cartesian. In Section 4.2 we derived the transformation that maps the components v_x, v_y, and v_z of a vector in Cartesian coordinates into its components v_r, v_θ, and v_φ in spherical coordinates:

$$\begin{pmatrix} v_r \\ v_\theta \\ v_\varphi \end{pmatrix} = \begin{pmatrix} \sin\theta\cos\varphi & \sin\theta\sin\varphi & \cos\theta \\ \cos\theta\cos\varphi & \cos\theta\sin\varphi & -\sin\theta \\ -\sin\varphi & \cos\varphi & 0 \end{pmatrix} \begin{pmatrix} v_x \\ v_y \\ v_z \end{pmatrix} . \tag{4.19}$$

This transformation is of the same form as the general transformation rule (22.45) for a tensor of rank one. However, the matrix of the coordinate transform now depends on the position; this is a result of the fact that the spherical coordinate system is curvilinear. When we take a derivative of the vector, the dependence of the transformation matrix on the position gives additional terms. For this reason tensor calculus for tensors in curvilinear coordinate systems is significantly more complex than tensor calculus for Cartesian tensors.

22.8 Products of tensors

One can form products of vectors in different ways. You have probably seen the inner product and outer product of vectors. There are also other ways in which one can make products of tensors. The first product of tensors that we consider is the *contraction*. You have already seen examples of this: the inner product of two vectors is one:

$$(\mathbf{u} \cdot \mathbf{v}) = u_i v_i . \tag{22.57}$$

Another example is matrix–vector multiplication:

$$(\mathbf{M} \cdot \mathbf{v})_i = M_{ij} v_j . \tag{22.58}$$

Note that according to the summation convention a summation over i is implied in (22.57) and a summation over j in (22.58).

These operations can be extended. Let $U_{i_1 i_2 \cdots i_n}$ be a tensor of rank n and $V_{j_1, j_2 \cdots j_m}$ a tensor of rank m. The contraction of **U** and **V** is defined by setting the last index of **U** equal to the first index of **V** and summing over this index:

$$(\mathbf{U} \cdot \mathbf{V})_{i_1 i_2 \cdots i_{n-1} k_2 \cdots k_m} = U_{i_1 \cdots i_{n-1} r} V_{r k_2 \cdots k_m} . \tag{22.59}$$

Note that we sum over the index r.

Problem a Show that the rank of this tensor is $n + m - 2$.

Problem b Show that the inner product of two vectors and matrix-vector multiplication are special cases of the contraction (22.59). Make sure you define n and m in each example. Verify that the rank of the result is indeed $n + m - 2$.

One can also apply a multiple contraction in which one sums over two or more indices of the tensors that one contracts. For example, the double contraction is given by

$$(\mathbf{U} : \mathbf{V})_{i_1 i_2 \cdots i_{n-2} k_3 \cdots k_m} = U_{i_1 \cdots i_{n-2} rs} V_{srk_3 \cdots k_m}. \tag{22.60}$$

Problem c Show that the resulting tensor is of rank $n + m - 4$.

A very important property is that *the contraction of two tensors is also a tensor.* The proof of this property is not difficult, it is mostly a tedious bookkeeping exercise, and the details can be found for example in references [24] and [87]. Here we will study some examples.

Problem d Show that the double contraction of a matrix **A** with the identity matrix is equal to the trace of **A** as defined in (22.36):

$$(\mathbf{A} : \mathbf{I}) = A_{ii} = \text{tr}\, \mathbf{A}. \tag{22.61}$$

The fact that the contraction of a tensor is also a tensor implies that the trace of a matrix is a tensor of rank zero: the trace of a matrix is a scalar. Therefore the trace of a matrix is invariant to unitary coordinate transformations.

Problem e There is one coordinate transformation in which the new basis vectors are aligned with the eigenvectors of **A**. Show that in that particular coordinate system the trace is equal to the sum of the eigenvectors λ_i. Use this result to derive that in any coordinate system the trace of a matrix is equal to the sum of the eigenvalues:

$$\text{tr}\, \mathbf{A} = \sum_i \lambda_i. \tag{22.62}$$

A matrix that is often used is the *Hessian* of a function f which is defined by the second partial derivatives of that function:

$$H_{ij} \equiv \frac{\partial^2 f}{\partial x_i \partial x_j}. \tag{22.63}$$

Problem e Show that **H** is a tensor of rank two by generalizing the derivation of problem e of Section 22.6 to the transformation properties of the second derivative.

Problem f Show that the trace of the Hessian is the Laplacian of f

$$\text{tr}\, H = \nabla^2 f. \tag{22.64}$$

The trace is invariant for unitary coordinate transformations and the Laplacian is the trace of a tensor; therefore the Laplacian is invariant for unitary coordinate transformations as well.

The second type of product of tensors is the direct product. The direct product of two tensors is formed by multiplying the different elements of the two tensors without carrying out a summation over repeated indices. An example of this is the dyad of two vectors that you encountered for example in Section 13.1 where the projection operator is written as the direct product of the unit vector $\hat{\mathbf{n}}$ with itself:

$$\mathbf{P} = \hat{\mathbf{n}}\hat{\mathbf{n}}^T. \tag{13.5}$$

In component form this expression is given by

$$P_{ij} = n_i n_j. \tag{22.65}$$

This idea can be generalized to form the direct product of a tensor **U** of rank n and a tensor **V** of rank m. This direct product is usually denoted with the symbol $\otimes$:

$$(\mathbf{U} \otimes \mathbf{V})_{i_1 \cdots i_n j_1 \cdots j_m} = U_{i_1 \cdots i_n} V_{j_1 \cdots j_m}. \tag{22.66}$$

Note that there are no repeated indices over which one sums.

Problem g Show that the rank of this direct product is $n + m$.

Problem h Show that the dyad (13.5) is a special example of the direct product (22.65).

Problem i The direct product of two tensors is also a tensor. Show this property by applying the transformation rule (22.49) to the tensors on the right-hand side of (22.66).

Problem j The gradient of a function is described in Chapter 5. The gradient $\mathbf{G} = \nabla \otimes \mathbf{v}$ of a vector $\mathbf{v}$ is defined as

$$G_{ij} \equiv \frac{\partial v_j}{\partial x_i}. \tag{22.67}$$

Show that this is a tensor of rank two.

22.9 Deformation and rotation again

As an important example of the direct product we consider in this section the strain tensor which measures the state of deformation in a medium. Consider a medium that is subject to some deformation and let us focus on two nearby points $\mathbf{r}$ and $\mathbf{r} + \delta\mathbf{r}$ in the medium as shown in Figure 22.6. During the deformation each point $\mathbf{r}$ is displaced by a vector $\mathbf{u}(\mathbf{r})$, which means that after the deformation the point $\mathbf{r}$ is located at $\mathbf{r} + \mathbf{u}(\mathbf{r})$ and the point $\mathbf{r} + \delta\mathbf{r}$ is located at $\mathbf{r} + \delta\mathbf{r} + \mathbf{u}(\mathbf{r} + \delta\mathbf{r})$. When the displacement of these neighboring points is the same ($\mathbf{u}(\mathbf{r} + \delta\mathbf{r}) = \mathbf{u}(\mathbf{r})$) the relative positions of these points are not changed. Therefore, the deformation is associated with the variation of the displacement vector $\mathbf{u}(\mathbf{r})$ with position $\mathbf{r}$. In order to describe the deformation we use the gradient tensor $\mathbf{D} = \nabla \otimes \mathbf{u}$ of the displacement vector.

$$D_{ij} \equiv \frac{\partial u_j}{\partial x_i}. \tag{22.68}$$

Problem a In a uniform translation ($\mathbf{u}(\mathbf{r}) = const.$) the medium is not deformed. Show that for this deformation the tensor $\mathbf{D}$ is indeed equal to zero.

The result from a previous problem reflects the fact that when the medium is not deformed the gradient tensor $\mathbf{D}$ is equal to zero. Unfortunately the reverse is not true; one may have a displacement that entails no deformation but the gradient tensor can be nonzero. We have seen in Section 22.3 an example that the displacement can be decomposed into a rotation and a deformation. Let us assume for the moment that the displacement is due to a rigid rotation around a point $\mathbf{r}_0$

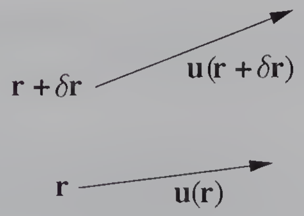

Fig. 22.6 The deformation of two nearby points $\mathbf{r}$ and $\mathbf{r} + \delta\mathbf{r}$

with rotation vector Ω. According to (13.24) the associated displacement is given by

$$\mathbf{u}^{rot}(\mathbf{r}) = \Omega \times (\mathbf{r} - \mathbf{r}_0). \tag{22.69}$$

Problem b Show that in component form this displacement is given by

$$\mathbf{u}^{rot}(\mathbf{r}) = \begin{pmatrix} \Omega_y(z - z_0) - \Omega_z(y - y_0) \\ \Omega_z(x - x_0) - \Omega_x(z - z_0) \\ \Omega_x(y - y_0) - \Omega_y(x - x_0) \end{pmatrix}. \tag{22.70}$$

Problem c Compute the partial derivatives of this vector to show that the associated gradient tensor is given by

$$\mathbf{D}^{rot} = \begin{pmatrix} 0 & \Omega_z & -\Omega_y \\ -\Omega_z & 0 & \Omega_x \\ \Omega_y & -\Omega_x & 0 \end{pmatrix}. \tag{22.71}$$

This is a remarkable result; although the displacement associated with the rotation depends on the position, the associated gradient tensor does not depend on the position. Note also that the gradient tensor does not depend on the point $\mathbf{r}_0$ around which the rotation takes place.

The gradient tensor $\mathbf{D}^{rot}$ is antisymmetric: each element is equal to the element on the other side of the main diagonal with an opposite sign: $D_{ij}^{rot} = -D_{ji}^{rot}$. Stated differently, the sum of each element and the element on the other side of the diagonal is equal to zero:

$$D_{ij}^{rot} + D_{ji}^{rot} = 0. \tag{22.72}$$

In general, any tensor of rank two can be written as the sum of a symmetric tensor and an antisymmetric tensor by using the following identity:

$$D_{ij} = \underbrace{\frac{1}{2}\left(D_{ij} + D_{ji}\right)}_{\text{deformation}} + \underbrace{\frac{1}{2}\left(D_{ij} - D_{ji}\right)}_{\text{rigid rotation}}. \tag{22.73}$$

Problem d Verify that this identity holds for any matrix $\mathbf{D}$. Show that the first term is symmetric and that the second term is anti-symmetric.

According to (22.72) the rotational component of the displacement does not contribute to the first term of (22.73). For this reason, the first term of this expression is used to characterize the deformation of the medium, while the second term characterizes the rotational component of the displacement. The strain tensor that characterizes the deformation of the medium ε is defined by the first term of (22.73).

Problem e Show that according to this definition

$$\varepsilon_{ij} = \frac{1}{2}\left(\frac{\partial u_j}{\partial x_i} + \frac{\partial u_i}{\partial x_j}\right). \tag{22.74}$$

Problem f We know that **D** is a tensor because it is the direct product of the gradient vector and the displacement vector, but we have to show that ε is a tensor. Do this by showing that when **D** is a tensor of rank two, then $\mathbf{D}^T$ also transforms as a tensor of rank 2. Then use that $\varepsilon = (1/2)\left(\mathbf{D} + \mathbf{D}^T\right)$ to show that ε is a tensor of rank two as well.

The strain tensor plays a crucial role in continuum mechanics because it is a measure of the degree of deformation in a medium. As shown in the tutorial of Lister and Williams [64] the partitioning between rotation and shear plays a vital role in structural geology since it is crucial in the generation of faults and shear zones.

22.10 Stress tensor

In general, when a medium is deformed, reaction forces are operative that tend to counteract the deformation. However, what do we mean by "reaction forces"? A force acts on something, but in a continuous medium there appears to be nothing to act on except a "point" in the medium. Since a point has zero mass this would lead to an infinite acceleration. This paradox can be resolved by considering a hypothetical cube in the medium as shown in Figure 22.7. The cube has six sides and we consider the force exerted by the rest of the medium on these six sides. Let

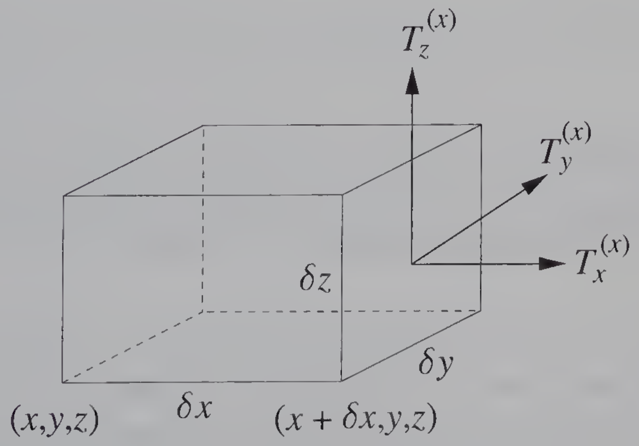

Fig. 22.7 The traction acting on a surface perpendicular to the x-direction.

us focus on the side on the right that is perpendicular to the x-direction. In other words the normal vector to this side is orientated in the x-direction.

The force on this surface depends on the size of the surface. A meaningful way to describe this is to use the *traction*, which is defined as the force per unit surface area. This traction $\mathbf{T}^{(x)}$ obviously has three components which are shown in Figure 22.7; these are denoted by $T_x^{(x)}$, $T_y^{(x)}$, and $T_z^{(x)}$. The subscript refers to the component of the traction, the superscript refers to the fact that this is the traction on a surface perpendicular to the x-direction. The component $T_x^{(x)}$ is normal to the surface and corresponds to a normal force, while the components $T_y^{(x)}$ and $T_z^{(x)}$ are parallel to the surface. The latter components are called the shear tractions because they cause shear motion of the medium.

When one considers the surface perpendicular to the y-direction one also finds a traction operating on that surface with components $T_x^{(y)}$, $T_y^{(y)}$, and $T_z^{(y)}$, and it is not clear at this point whether the traction $\mathbf{T}^{(y)}$ has any relation to the traction $\mathbf{T}^{(x)}$ that acts on a surface perpendicular to the x-direction. The tractions acting on the three surfaces can be grouped in a 3×3 matrix:

$$\sigma \equiv \begin{pmatrix} \vdots & \vdots & \vdots \\ \mathbf{T}^{(x)} & \mathbf{T}^{(y)} & \mathbf{T}^{(z)} \\ \vdots & \vdots & \vdots \end{pmatrix} = \begin{pmatrix} T_x^{(x)} & T_x^{(y)} & T_x^{(z)} \\ T_y^{(x)} & T_y^{(y)} & T_y^{(z)} \\ T_z^{(x)} & T_z^{(y)} & T_z^{(z)} \end{pmatrix}. \tag{22.75}$$

The quantity σ is called the *stress tensor.*

The stress tensor has a remarkable property that is reminiscent of the property of the gradient which was introduced in Section 5.1. The gradient is so useful because once one knows the three components of the gradient, one can compute the change of a function in *any* direction. The stress tensor gives the components of the traction to surfaces that are perpendicular to the coordinates axes. One can show that the traction on a surface that is perpendicular to an arbitrary unit vector $\hat{\mathbf{n}}$ is given by

$$\mathbf{T} = \sigma \cdot \hat{\mathbf{n}}. \tag{22.76}$$

Problem a The proof of this identity is actually very simple. Consult Section 16.3 of Butkov [24] or your favorite book on continuum mechanics or mathematical physics (that is, aside from this book) for a proof.

We have called σ the stress tensor, but we have not shown that σ is a tensor. However, one can show that when a quantity is contracted with a tensor and the result is also a tensor, then this quantity must be a tensor as well. In (22.76) we know that $\mathbf{T}$ is a tensor because it is a force (normalized with surface area) and we have shown in Section 22.6 that the force is a tensor of rank one. The normal vector $\hat{\mathbf{n}}$ is also a tensor of rank one because it transforms in the same way as the position vector. Hence we need to show from (22.76) and the fact that both $\mathbf{T}$ and $\hat{\mathbf{n}}$ are tensors of rank one that $\boldsymbol{\sigma}$ is a tensor as well.

In components (22.76) is written as

$$T_i = \sigma_{ij} n_j, \tag{22.77}$$

while the same expression in a transformed coordinate system is given by

$$T_i' = \sigma_{ij}' n_j'. \tag{22.78}$$

Before we determine the transformation property of the stress tensor $\boldsymbol{\sigma}$, we first express the original vector $\mathbf{v}$ in terms of the transformed vector $\mathbf{v}'$ using the transformation rule

$$v_i' = C_{ij} v_j. \tag{22.45}$$

Problem b Multiply this expression on the left by C_{ki}^{-1} and sum over i to obtain $C_{ki}^{-1} v_i' = C_{ki}^{-1} C_{ij} v_j$. Use the fact that $\mathbf{C}$ is unitary on the left-hand side and carry out the matrix multiplication on the right-hand side to show that $C_{ik} v_i' = v_k$. Finally rename $i \to j$ and $k \to i$ to obtain

$$v_i = C_{ji} v_j'. \tag{22.79}$$

It is interesting to compare this expression for the original vector $\mathbf{v}$ in terms of the transformed vector $\mathbf{v}'$ with (22.45) which gives the transformed vector given the original vector. The only difference is the order of the subscripts in the coordinate transform $\mathbf{C}$. This is due to the fact that this matrix is unitary so that inversion amounts to interchanging the indices: $C_{ij}^{-1} = C_{ij}^T = C_{ji}$.

Problem c Insert (22.79) into (22.77) in order to express the unprimed vectors $\mathbf{T}$ and $\hat{\mathbf{n}}$ in terms of their transformed vectors to obtain

$$C_{ki} T_k' = \sigma_{ij} C_{lj} n_l'. \tag{22.80}$$

Problem d Write $C_{ki} = C_{ik}^T = C_{ik}^{-1}$ on the left-hand side and multiply by C_{mi} to obtain

$$T_m' = C_{mi} C_{lj} \sigma_{ij} n_l'. \tag{22.81}$$

Problem e We want to compare this expression with (22.78) in order to find $\boldsymbol{\sigma}'$. However, the indices are different. As we noted earlier, the names of indices are irrelevant. Rename the indices in (22.81) so that this expression can be compared directly with (22.78) and use this to show that

$$\sigma'_{ij} = C_{ir}C_{js}\sigma_{rs}. \tag{22.82}$$

This is just the transformation rule (22.49) for a tensor of rank two. We have thus shown that the stress tensor is indeed a tensor of rank two. This is due to the fact that when an object is contracted with a tensor to give another tensor, this object must be a tensor as well. Using this property we can use a bootstrap procedure to find higher order tensors. An important example of this is the elasticity tensor which relates the stress to the strain:

$$\sigma_{ij} = c_{ijkl}\varepsilon_{kl}. \tag{22.83}$$

This expression generalizes the elastic force $F = -kx$ in a spring to continuous media; it is known as Hooke's law. The quantity c_{ijkl} must be a tensor because we know that both the stress σ and the strain ε are tensors. This means that c_{ijkl} is a tensor of rank four.

22.11 Why pressure in a fluid is isotropic

Finally we have reached the point where we can use tensors to learn about physics. As a first example we consider the pressure in a fluid (or gas). The pressure is the normal force per unit surface area. It is an observational fact that in a gas or fluid this force does not depend on the *orientation* of this surface. (This is why the weather forecaster speaks about a pressure of 1020 mbar rather than saying that the pressure is 1015 mbar in the vertical direction and 1025 mbar in the horizontal direction.) In this section we discover why the pressure is independent of direction.

In order to understand this we return to the stress tensor (22.75) and we consider the traction acting on a surface perpendicular to the x-direction. As shown in Figure 22.7, $T_x^{(x)}$ gives the traction normal to this surface while $T_y^{(x)}$ and $T_z^{(x)}$ give the shear traction that acts on this surface. This reasoning can be used for all the surfaces; the diagonal elements $T_x^{(x)}$, $T_y^{(y)}$, and $T_z^{(z)}$ of the stress tensor give the normal tractions while all the other elements give the shear tractions. In a fluid, there are no shear tractions because a fluid has zero shear strength. This means that in a fluid the stress tensor is diagonal:

$$\sigma = -\begin{pmatrix} p_x & 0 & 0 \\ 0 & p_y & 0 \\ 0 & 0 & p_z \end{pmatrix}. \tag{22.84}$$

The diagonal elements p_i denote the pressure in the three directions. The minus sign reflects the fact that a positive pressure corresponds to a force that is directed *inwards*. In this section we show that the pressure is isotropic, in other words that the diagonal elements are identical.

Let us see what happens to the stress tensor (22.84) when we rotate the coordinate system through 45 degrees around the z-axis. For a rotation in two dimensions the rotation is given by (22.23). Setting the rotation angle φ to 45 degrees and extending the result to three dimensions gives the following matrix representation of this coordinate transformation:

$$\mathbf{C} = \begin{pmatrix} 1/\sqrt{2} & -1/\sqrt{2} & 0 \\ 1/\sqrt{2} & 1/\sqrt{2} & 0 \\ 0 & 0 & 1 \end{pmatrix}. \tag{22.85}$$

Problem a Verify that this coordinate transformation is unitary.

Problem b Use the transformation property of a tensor of rank two that the stress tensor (22.84) in the rotated coordinate system is given by

$$\boldsymbol{\sigma}' = -\begin{pmatrix} \left(p_x + p_y\right)/2 & \left(p_x - p_y\right)/2 & 0 \\ \left(p_x - p_y\right)/2 & \left(p_x + p_y\right)/2 & 0 \\ 0 & 0 & p_z \end{pmatrix}. \tag{22.86}$$

In a fluid, the stress tensor is diagonal in *any* coordinate system since the shear tractions vanish in any coordinate system. This means that the off-diagonal elements of $\boldsymbol{\sigma}'$ must be equal to zero, hence $p_x = p_y$.

Problem c Find a suitable coordinate transform to show that $p_x = p_z$.

This means that all the diagonal elements are identical; this quantity is referred to as the pressure: $p_x = p_y = p_z = p$. In an acoustic medium such as a fluid or gas the stress tensor is therefore given by

$$\boldsymbol{\sigma} = -p\,\mathbf{I}, \tag{22.87}$$

so that pressure is indeed independent of direction.

Note that we have done something truly remarkable; we have derived a physical law (“the pressure is isotropic”) from the invariance of a property (“the shear stress is zero in a fluid”) under a coordinate transformation.

22.12 Special relativity

One of the most spectacular applications of tensor calculus is the theory of relativity which describes the physics of objects and fields at very high speeds. The theory of general relativity accounts for the fact that mass in the universe leads to a non-Cartesian structure of space-time [76]; by definition this cannot be treated with the Cartesian tensors used in this chapter. The theory of special relativity describes how different observers who both use Cartesian coordinate systems that move at great speeds with respect to each other describe the same physical phenomena. A clear physical description of the theory of special relativity is given by Taylor and Wheeler [105]. In this section we use the notation used by Muirhead [73] who uses a complex time variable.

Central to the theory of special relativity is the notion that space and time are intricately linked. The three position variables x, y, and z as well as time t are placed in a four-dimensional vector, called a four-vector:

$$\mathbf{x} = \begin{pmatrix} x \\ y \\ z \\ ict \end{pmatrix} . \tag{22.88}$$

In this expression c is the speed of light and $i = \sqrt{-1}$. The fact that the last component is complex leads to the surprising result that the length of the four-vector can be negative because

$$|\mathbf{x}|^2 = (\mathbf{x} \cdot \mathbf{x}) = x^2 + y^2 + z^2 - c^2t^2 \tag{22.89}$$

and there is no reason why the last term cannot dominate the other terms.

Suppose we have one observer who uses unprimed variables, and suppose that another observer moves in the x-direction with a relative velocity v with respect to the first observer. The two coordinate systems of the observers are then related by a *Lorentz transformation* [73]:

$$\mathbf{L} = \begin{pmatrix} 1/\sqrt{1-\frac{v^2}{c^2}} & 0 & 0 & i\frac{v}{c}/\sqrt{1-\frac{v^2}{c^2}} \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ -i\frac{v}{c}/\sqrt{1-\frac{v^2}{c^2}} & 0 & 0 & 1/\sqrt{1-\frac{v^2}{c^2}} \end{pmatrix} . \tag{22.90}$$

Problem a Show that the Lorentz transform is unitary by showing that $\mathbf{L}^T\mathbf{L} = \mathbf{I}$.

Note that the transpose $\mathbf{L}^T$ is used here and *not* the Hermitian conjugate $\mathbf{L}^\dagger$ that is defined as the transpose and the complex conjugate: $L^\dagger_{ij} \equiv L^*_{ji}$. The theory of special relativity states that the four-vector **x** is a tensor of rank one. This implies that $|\mathbf{x}|^2$ is a scalar, which means that both observers will assign the same value to this property, so that

$$x'^2 + y'^2 + z'^2 - c^2 t'^2 = x^2 + y^2 + z^2 - c^2 t^2. \tag{22.91}$$

Problem b Use the fact that **x** is a tensor of rank one to show that after a Lorentz transformation the x'- and t'-coordinates are given by

$$\left.\begin{aligned} x' &= (x - vt) \Big/ \sqrt{1 - \frac{v^2}{c^2}}, \\ t' &= \left(t - \frac{vx}{c^2}\right) \Big/ \sqrt{1 - \frac{v^2}{c^2}}. \end{aligned}\right\} \tag{22.92}$$

Problem c Verify that this transformation (together with $y = y'$ and $z = z'$) indeed satisfies the identity (22.91).

The transformation (22.92) means that space and time are mixed in a Lorentz transformation. The distinction between space and time largely disappears in the theory of relativity! The term $\left(1 - v^2/c^2\right)^{-1/2}$ on the right-hand side leads to the clocks of the two observers running at different speeds, contracting rods and other phenomena that are counterintuitive but which have been confirmed experimentally [105]. Note that the contraction terms $\left(1 - v^2/c^2\right)^{-1/2}$ are crucial in problem c in establishing that (22.91) is satisfied.

The theory of special relativity has many surprises. We have already seen in Section 22.7 that a magnetic field does not behave as a tensor of rank one under coordinate transformations in three dimensions. In fact, when we also consider coordinate transformations between moving coordinate systems the electric field and the magnetic fields are mixed! According to the theory of special relativity [73] the magnetic field **B** and the electric field **E** transform as the following tensor of rank two:

$$\mathbf{F} = \begin{pmatrix} 0 & B_z & -B_y & -iE_x/c \\ -B_z & 0 & B_x & -iE_y/c \\ B_y & -B_x & 0 & -iE_z/c \\ iE_x/c & iE_y/c & iE_z/c & 0 \end{pmatrix}. \tag{22.93}$$

The momentum of a particle is also given by a four-vector; this momentum-energy vector [73] is given by

$$\mathbf{p} = \begin{pmatrix} p_x \\ p_y \\ p_z \\ iE/c \end{pmatrix}, \tag{22.94}$$

where E is the energy of the particle. This four-vector transforms as a tensor of rank one under a Lorentz transformation [73].

Problem d Use the property that $\mathbf{p}$ is a tensor of rank one to show that $p^2 - E^2/c^2$ is invariant, where $p^2 = p_x^2 + p_y^2 + p_z^2$.

Problem e In general the energy of a body is given by the energy of that body at rest plus a contribution due to the motion of the body. This rest energy is denoted by E_0. Show that for the particle at rest the norm of the four-vector $\mathbf{p}$ is given by $|\mathbf{p}|^2 = -E_0^2/c^2$ and use the result of problem d to show that

$$E = \sqrt{E_0^2 + p^2c^2}. \tag{22.95}$$

This expression holds for any value of the momentum. For the moment we consider a particle that moves much slower than the speed of light ($p = mv \ll mc$). We shall show that for such a particle the rest energy E_0 is much larger than the energy pc that is due to the motion.

Problem f Make a Taylor series expansion of (22.95) in the parameter pc/E_0 to show that for slow speeds

$$E = E_0 + \frac{c^2p^2}{2E_0} - \frac{c^4p^4}{8E_0^3} + \cdots. \tag{22.96}$$

The first term on the right-hand side corresponds to the rest energy which we do not yet know. The second term is quadratic in the momentum. For small velocities the laws of classical mechanics hold and the kinetic energy is given by $\frac{1}{2}m_0v^2$. Using the classical relation $p = m_0v$, the kinetic energy is in classical mechanics given by $p^2/2m_0$. In classical mechanics the terms of order $(cp/E_0)^4$ are ignored. The kinetic energy $p^2/2m_0$ in classical mechanics is therefore equal to the second term on the right-hand side of (22.96).

Problem g Derive from this statement that

$$E_0 = m_0c^2. \tag{22.97}$$

You have just derived the famous relation that relates the rest energy of a particle to its rest mass. Note that the only ingredients that you have used were the fact that the four-vector **p** transforms like a tensor of rank one under coordinate transforms and some elementary results from classical mechanics! The implications of (22.97) are profound. It reflects the fact that matter can be seen as a condensed form of energy. In nuclear reactions this is used for the benefit (or demise) of mankind because the mass of the end-result of some nuclear reactions is smaller than the mass of the ingredients for that reaction. This mass difference is released in the form of radiation and the kinetic energy (heat) of the resulting particles.

Problem h In the derivation of (22.97) we assumed that the rest energy in the classical limit is much larger than the kinetic energy. Show that the ratio of the second term on the right-hand side of (22.96) to the rest mass is given by

$$\frac{c^2 p^2}{2E_0}/E_0 = \frac{1}{2}(v/c)^2 . \tag{22.98}$$

Problem i The velocity with which a rocket can overcome the attraction of the Earth is called the *escape velocity* [55]; it has the numerical value of 11 184 km/s. Compute the ratio v/c for a rocket that leaves the Earth at the escape velocity and compute the ratio of the third term to the second term on the right-hand side of (22.96).

The third term on the right-hand side of (22.96) is a relativistic correction term that gives the leading order correction to the classical kinetic energy.

At this point you might think that the theory only has implications for high-velocity bodies in cosmological problems. This is, however, not true; the electrons in atoms move so fast that relativistic effects leave an imprint on microscopic bodies as well. For example, the relativistic correction term $c^4 p^4/8E_0^3$ in (22.96) leads in quantum theory to a measurable shift in the frequency of light emitted by excited hydrogen atoms that is called *fine-structure splitting* [91]. The Global Positioning System is a navigation system that utilizes the radio signals emitted from a number of satellites that move around the Earth at a velocity that is much smaller than the speed of light. Relativistic effects cause the clocks of these satellites to run at a slightly different rate than earthbound clocks. This is a tiny effect, but the extreme time-accuracy that is required for the precision of navigation that is required makes it necessary to correct for the change in the clock rate due to relativistic effects [6].

23

Perturbation theory

From this book and most other books on mathematical physics you may have obtained the impression that most equations in the physical sciences can be solved. This is actually not true; most textbooks (including this book) give an unrepresentative state of affairs by only showing the problems that *can* be solved in closed form. It is an interesting paradox that as our theories of the physical world become more accurate, the resulting equations become more difficult to solve. In classical mechanics the problem of two particles that interact with a central force can be solved in closed form, but the three-body problem in which three particles interact has no analytical solution. In quantum mechanics, the one-body problem of a particle that moves in a potential can be solved for a limited number of situations only: for the free particle, the particle in a box, the harmonic oscillator, and the hydrogen atom. In this sense the one-body problem in quantum mechanics has no general solution. This shows that as a theory becomes more accurate, the resulting complexity of the equations makes it often more difficult to actually find solutions.

One way to proceed is to compute numerical solutions of the equations. Computers are a powerful tool and can be extremely useful in solving physical problems. Another approach is to find approximate solutions to the equations. In Chapter 12, scale analysis was used to drop from the equations terms that appear to be irrelevant. In this chapter, a systematic method is introduced to account for terms in the equations that are small but that make the equations difficult to solve. The idea is that a complex problem is compared to a simpler problem that can be solved in closed form, and to consider these small terms as a perturbation to the original equation. The theory of this chapter then makes it possible to determine how the solution is perturbed by the perturbation in the original equation; this technique is called *perturbation theory*. A classic reference on perturbation theory has been written by Nayfeh [74]. The book by Bender and Orszag [14] gives a useful and illustrative overview of a wide variety of perturbation methods.

The central idea of perturbation theory is introduced for an algebraic equation in Section 23.1. Sections 23.2, 23.3, and 23.5 contain important applications of perturbation theory to differential equations. As shown in Section 23.4, perturbation theory has a limited domain of applicability, and this may depend on the way the perturbation problem is formulated. Finally, it is shown in Section 23.7 that not every perturbation problem is well behaved; this leads to singular perturbation theory. Chapter 24 is devoted to the asymptotic evaluation of integrals.

23.1 Regular perturbation theory

As an introduction to perturbation theory let us consider the following equation

$$x^3 - 4x^2 + 4x = 0.01. \tag{23.1}$$

Let us for the moment assume that we do not know how to find the roots of a third order polynomial, so we cannot solve this equation. The problem is the small term 0.01 on the right-hand side. If this term were equal to zero, the resulting equation can be solved; $x^3 - 4x^2 + 4x = 0$ is equivalent to $x(x^2 - 4x + 4) = x(x-2)^2 = 0$, which has the solutions $x = 0$ and $x = 2$. In Figure 23.1 the polynomial of (23.1) is shown by the thick solid line; it is indeed equal to zero for $x = 0$ and $x = 2$.

The problem that we face is that the right-hand side of (23.1) is *not* equal to zero. In perturbation theory one studies the perturbation of the solution under a perturbation of the original equation. In order to do this, we replace the original equation (23.1) by the more general equation

$$x^3 - 4x^2 + 4x = \varepsilon. \tag{23.2}$$

When $\varepsilon = 0.01$ this equation is identical to the original problem, while for $\varepsilon = 0$ it reduces to the unperturbed problem that we can solve in closed form. It may appear

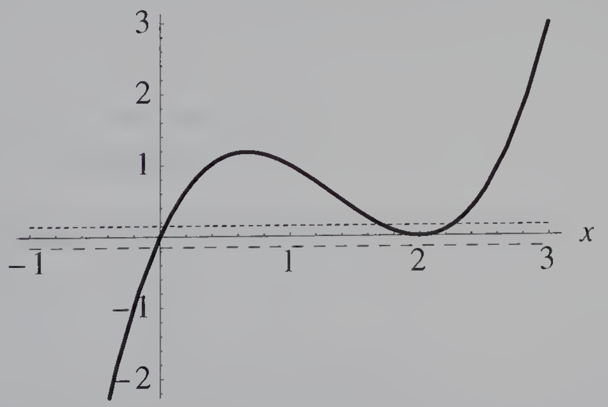

Fig. 23.1 The polynomial $x^3 - 4x^2 + 4x$ (thick solid line) and the lines $\varepsilon = 0.15$ (dotted line) and $\varepsilon = -0.15$ (dashed line).

that we have made the problem more complex because we still need to solve the same equation as our original equation, but it now contains a new variable ε as well! However, this is also the strength of this approach.

The solution of (23.2) is a function of ε so that

$$x = x(\varepsilon)\ . \tag{23.3}$$

In Section 3.1 the Taylor series was used to approximate a function $f(x)$ by a power series in the variable x:

$$f(x) = f(0) + x\frac{df}{dx}(x=0) + \frac{x^2}{2!}\frac{d^2 f}{dx^2}(x=0) + \cdots . \tag{3.11}$$

When the solution x of (23.2) depends in a regular way on ε, this solution can also be written as a similar power series by making the substitutions $x \rightarrow \varepsilon$ and $f \rightarrow x$ in (3.11):

$$x(\varepsilon) = x(0) + \varepsilon\frac{dx}{d\varepsilon}(\varepsilon=0) + \frac{\varepsilon^2}{2!}\frac{d^2 x}{d\varepsilon^2}(\varepsilon=0) + \cdots . \tag{23.4}$$

This expression is not very useful because we need the derivative $dx/d\varepsilon$ and higher derivatives $d^n x/d\varepsilon^n$ as well; in order to compute these derivatives we need to find the solution $x(\varepsilon)$ first, but this is just what we are trying to do. There is, however, another way to determine the series (23.4). Let us write the solution $x(\varepsilon)$ as a power series in ε

$$x(\varepsilon) = x_0 + \varepsilon x_1 + \varepsilon^2 x_2 + \cdots . \tag{23.5}$$

The coefficients x_n are not known at this point, but once we know them the solution x can be found by inserting the numerical value $\varepsilon = 0.01$. In practice one truncates the series (23.5); it is this truncation that makes perturbation theory an approximation.

When the series (23.5) is inserted into (23.2) one needs to compute x^2 and x^3 when x is given by (23.5). Let us first consider the x^2-term. The square of a sum of terms is given by

$$\begin{aligned}(a+b+c+\cdots)^2 &= a^2+b^2+c^2+\cdots \\ &\quad + 2ab + 2ac + 2bc + \cdots .\end{aligned} \tag{23.6}$$

Let us apply this to the series (23.5) and retain only the terms up to order ε^2, this gives

$$\begin{aligned}\left(x_0 + \varepsilon x_1 + \varepsilon^2 x_2 + \cdots\right)^2 &= x_0^2 + \varepsilon^2 x_1^2 + \varepsilon^4 x_2^2 + \cdots \\ &\quad + 2\varepsilon x_0 x_2 + 2\varepsilon^2 x_0 x_2 + 2\varepsilon^3 x_1 x_2 + \cdots .\end{aligned} \tag{23.7}$$

If we are only are interested in retaining the terms up to order ε^2, the terms $\varepsilon^4 x_2^2$ and $2\varepsilon^3 x_1 x_2$ in this expression can be ignored. Collecting terms of equal powers of

ε then gives

$$\left(x_0+\varepsilon x_1+\varepsilon^2 x_2+\cdots\right)^2=x_0^2+2\varepsilon x_0 x_1+\varepsilon^2\left(x_1^2+2x_0x_2\right)+O(\varepsilon^3). \quad (23.8)$$

A similar expansion in powers of ε can be used for the term x^3. This expansion is based on the identity

$$\begin{aligned}(a+b+c+\cdots)^3&=a^3+b^3+c^3+\cdots\\&\quad+3a^2b+3ab^2+3a^2c+3ac^2+3b^2c+3bc^2+\cdots.\end{aligned} \quad (23.9)$$

Problem a Apply this identity to the series (23.5), collect together all the terms with equal powers of ε and show that up to order ε^2 the result is given by

$$\left(x_0+\varepsilon x_1+\varepsilon^2 x_2+\cdots\right)^3=x_0^3+3\varepsilon x_0^2 x_1+3\varepsilon^2\left(x_0x_1^2+x_0^2x_2\right)+O(\varepsilon^3). \quad (23.10)$$

Problem b At this point we can express all the terms in (23.2) in a power series of ε. Insert (23.5), (23.8), and (23.10) into the original equation (23.2) and collect together terms of equal powers of ε to derive that

$$\begin{aligned}&x_0^3-4x_0^2+4x_0\\&\quad+\varepsilon\left(3x_0^2x_1-8x_0x_1+4x_1-1\right)\\&\quad+\varepsilon^2\left(3x_0x_1^2+x_0^2x_2-4x_1^2-8x_0x_2+4x_2\right)+\cdots=0.\end{aligned} \quad (23.11)$$

In this and subsequent expressions the dots denote terms of order $O(\varepsilon^3)$. The term -1 in the term that multiplies ε comes from the right-hand side of (23.2).

At this point we use that ε does not have a fixed value, but that it can take any value within certain bounds. This means that expression (23.11) must be satisfied for a range of values of ε. This can only be the case when the coefficients that multiply the different powers ε^n are equal to zero. This means that (23.11) is equivalent to the following system of equations which consists of the terms that multiply the terms ε^0, ε^1 and ε^2 respectively:

$$\left.\begin{array}{ll}O(1)\text{-terms:} & x_0^3-4x_0^2+4x_0=0,\\ O(\varepsilon)\text{-terms:} & 3x_0^2x_1-8x_0x_1+4x_1-1=0,\\ O(\varepsilon^2)\text{-terms:} & 3x_0x_1^2+x_0^2x_2-4x_1^2-8x_0x_2+4x_2=0.\end{array}\right\} \quad (23.12)$$

You may wonder whether we have not made the problem more complex. We started with a single equation for a single variable x, and now we have a system of coupled equations for many variables. However, we could not solve (23.2) for the single variable x, while it is not difficult to solve (23.12).

Problem c Show that (23.12) can be rewritten in the following form:

$$\left.\begin{aligned} &x_0^3 - 4x_0^2 + 4x_0 = 0 , \\ &\left(3x_0^2 - 8x_0 + 4\right) x_1 = 1 , \\ &\left(x_0^2 - 8x_0 + 4\right) x_2 = (4 - 3x_0)\, x_1^2 \, . \end{aligned}\right\} \tag{23.13}$$

The first equation is simply the unperturbed problem, this has the solutions $x_0 = 0$ and $x_0 = 2$. For reasons that will become clear in Section 23.7 we focus here on the solution $x_0 = 0$ only. Given x_0, the parameter x_1 follows from the second equation because this is a linear equation in x_1. The last equation is a linear equation in the unknown x_2 which can easily be solved once x_0 and x_1 are known.

Problem d Solve (23.13) in this way to show that the solution near $x = 0$ is given by

$$x_0 = 0, \qquad x_1 = \frac{1}{4}, \qquad x_2 = \frac{1}{16} \, . \tag{23.14}$$

Now we are close to the final solution of our problem. The coefficients of the previous expression can be inserted into the perturbation series (23.5) so that the solution as a function of ε is given by

$$x = 0 + \frac{1}{4}\varepsilon + \frac{1}{16}\varepsilon^2 + O(\varepsilon^3) \, . \tag{23.15}$$

At this point we can revert to the original equation (23.1) by inserting the numerical value $\varepsilon = 0.01$, which gives:

$$x = \frac{1}{4} \times 10^{-2} + \frac{1}{16} \times 10^{-4} + O(10^{-6}) = 0.002506 \, . \tag{23.16}$$

It should be noted that this is an approximate solution because the terms of order ε^3 and higher have been ignored. This is indicated by the term $O(10^{-6})$ in (23.16). Assuming that the error made by truncating the perturbation series is of the same order as the first term that is truncated, the error in the solution (23.16) is of the order 10^{-6}. For this reason the number on the right-hand side of (23.16) is given to six decimals; the last decimal is of the same order as the truncation error.

If this result is not sufficiently accurate for the application that one has in mind, then one can easily extend the analysis to higher powers ε^n in order to reduce the truncation error of the truncated perturbation series. Although the algebra resulting from doing this can be tedious, there is no reason why this analysis cannot be extended to higher orders.

A truly formal analysis of perturbation problems can be difficult. For example, the perturbation series (23.5) converges only for sufficiently small values of ε. It is often not clear whether the employed value of ε (in this case $\varepsilon = 0.01$) is sufficiently small to ensure convergence. Even when a perturbation series does not converge for a given value of ε, one can often obtain a useful approximation to the solution by truncating the perturbation series at a suitably chosen order [14]. In this case one speaks of an *asymptotic series.*

When one has obtained an approximate solution of a perturbation problem, one can sometimes substitute it back into the original equation to verify whether this solution indeed satisfies the equation with an acceptable accuracy. For example, inserting the numerical value $x = 0.002\,506$ in (23.1) gives

$$x^3 - 4x^2 + 4x = 0.009\,998\,9 = 0.01 - 0.000\,001\,1\ . \tag{23.17}$$

This means that the approximate solution satisfies (23.1) with a *relative* error that is given by $0.000\,001\,1/0.01 = 10^{-4}$. This is a very accurate result given the fact that only three terms were retained in the perturbation analysis of this section.

23.2 Born approximation

In many scattering problems one wants to account for the scattering of waves by the heterogeneities in the medium. Usually these problems are so complex that they cannot be solved in closed form. Suppose one has a background medium in which scatterers are embedded. When the background medium is sufficiently simple, one can solve the wave propagation problem for this background medium. For example, in Section 19.3 we computed the Green's function for the Helmholtz equation in a homogeneous medium.

In this section we consider the Helmholtz equation with a *variable* velocity $c(\mathbf{r})$ as an example of the application of perturbation theory to scattering problems. This means we consider the wave field $p(\mathbf{r}, \omega)$ in the frequency domain that satisfies the following equation:

$$\nabla^2 p(\mathbf{r}, \omega) + \frac{\omega^2}{c^2(\mathbf{r})} p(\mathbf{r}, \omega) = S(\mathbf{r}, \omega)\ . \tag{23.18}$$

In this expression $S(\mathbf{r}, \omega)$ denotes the source that generates the wave field. In order to facilitate a systematic perturbation analysis we decompose $1/c^2(\mathbf{r})$ into a term $1/c_0^2$ that accounts for a homogeneous reference model and a perturbation:

$$\frac{1}{c^2(\mathbf{r})} = \frac{1}{c_0^2}\left[1 + \varepsilon n(\mathbf{r})\right] . \tag{23.19}$$

In this expression ε is a small parameter which measures the strength of the heterogeneity. The function $n(\mathbf{r})$ gives the spatial distribution of the heterogeneity. Combining the previous expressions it follows that the wave field satisfies the following expression:

$$\nabla^2 p(\mathbf{r},\omega)+\frac{\omega^2}{c_0^2}\left[1+\varepsilon n(\mathbf{r})\right]\ p(\mathbf{r},\omega)=S(\mathbf{r},\omega). \tag{23.20}$$

The solution $p(\mathbf{r},\omega)$ of this expression is a function of the scattering strength ε; for sufficiently small values of ε it can be written as a power series in ε:

$$p=p_0+\varepsilon p_1+\varepsilon^2 p_2+\cdots. \tag{23.21}$$

Problem a Insert the perturbation series (23.21) into (23.20), collect together the terms that multiply equal powers of ε and show that the terms that multiply the different powers of ε give the following equations:

$$\left.\begin{array}{ll}
O(1): & \nabla^2 p_0(\mathbf{r},\omega)+\dfrac{\omega^2}{c_0^2}p_0(\mathbf{r},\omega)=S(\mathbf{r},\omega),\\
O(\varepsilon): & \nabla^2 p_1(\mathbf{r},\omega)+\dfrac{\omega^2}{c_0^2}p_1(\mathbf{r},\omega)=-\dfrac{\omega^2}{c_0^2}n(\mathbf{r})p_0(\mathbf{r},\omega),\\
O(\varepsilon^2): & \nabla^2 p_2(\mathbf{r},\omega)+\dfrac{\omega^2}{c_0^2}p_2(\mathbf{r},\omega)=-\dfrac{\omega^2}{c_0^2}n(\mathbf{r})p_1(\mathbf{r},\omega),\\
 & \vdots
\end{array}\right\} \tag{23.22}$$

The first expression gives the Helmholtz equation for a homogeneous medium. The source of the unperturbed wave p_0 is given by $S(\mathbf{r},\omega)$ which is also the source of the perturbed problem. The source of the first order perturbation p_1 is given by the right-hand side of the second equation, hence the source of p_1 is given by $-\left(\omega^2/c_0^2\right)n(\mathbf{r})p_0(\mathbf{r},\omega)$. This means that the source of p_1 is proportional to the inhomogeneity $n(\mathbf{r})$ of the medium. Physically this corresponds to the fact that the heterogeneity is the source of the scattered waves. The source of p_1 is also proportional to the unperturbed wave field p_0. The reason for this is that the generation of the scattered waves depends on the local perturbation of the medium as well as on the strength of the wave field at the location of the scatterers.

Each of the equations in (23.22) is of the form $\nabla^2 p(\mathbf{r},\omega)+\omega^2/c_0^2 p(\mathbf{r},\omega)=F(\mathbf{r},\omega)$. According to the theory of Section 18.4 the solution to this expression is given by

$$p(\mathbf{r},\omega)=\int G_0(\mathbf{r},\mathbf{r}';\omega)F(\mathbf{r}',\omega)\,dV', \tag{23.23}$$

where the unperturbed Green's function $G_0(\mathbf{r},\mathbf{r}';\omega)$ is the response in a homogeneous medium at location $\mathbf{r}$ due to a point source at location $\mathbf{r}'$:

$$\nabla^2 G_0(\mathbf{r},\mathbf{r}';\omega)+\frac{\omega^2}{c_0^2}G_0(\mathbf{r},\mathbf{r}';\omega)=\delta(\mathbf{r}-\mathbf{r}'). \tag{23.24}$$

The specific form of the unperturbed Green's function in one, two, and three dimensions is given in (19.43). From this point on it is not shown explicitly that the solution and the Green's function depend on the angular frequency ω, but it should be kept in mind that all the results in this section depend on frequency.

Problem c Use these results to show that the solution of (23.22) is given by

$$\left.\begin{aligned} p_0(\mathbf{r}) &= \int G_0(\mathbf{r},\mathbf{r}')S(\mathbf{r}')\,dV', \\ p_1(\mathbf{r}) &= -\frac{\omega^2}{c_0^2}\int G_0(\mathbf{r},\mathbf{r}')n(\mathbf{r}')p_0(\mathbf{r}')\,dV', \\ p_2(\mathbf{r}) &= -\frac{\omega^2}{c_0^2}\int G_0(\mathbf{r},\mathbf{r}')n(\mathbf{r}')p_1(\mathbf{r}')\,dV', \\ &\vdots \end{aligned}\right\} \tag{23.25}$$

Problem d Insert the expression for the unperturbed wave p_0 into the second equation (23.25) to show that the first order perturbation is given by

$$p_1(\mathbf{r}) = -\frac{\omega^2}{c_0^2}\iint G_0(\mathbf{r},\mathbf{r}_1)n(\mathbf{r}_1)G_0(\mathbf{r}_1,\mathbf{r}_0)S(\mathbf{r}_0)\,dV_1 dV_0. \tag{23.26}$$

Note that the integration variable $\mathbf{r}'$ has been relabeled as $\mathbf{r}_0$ and $\mathbf{r}_1$ respectively.

Problem e Insert this result in the last line of (23.25) to derive that the second order perturbation is given by

$$p_2(\mathbf{r}) = \frac{\omega^4}{c_0^4} \times \iiint G_0(\mathbf{r},\mathbf{r}_2)n(\mathbf{r}_2)G_0(\mathbf{r}_2,\mathbf{r}_1)n(\mathbf{r}_1)G_0(\mathbf{r}_1,\mathbf{r}_0)S(\mathbf{r}_0)\,dV_2 dV_1 dV_0. \tag{23.27}$$

Inserting this result and (23.25) in the perturbation series (23.21) finally gives the following perturbation series for the scattered waves:

$$\begin{aligned} p(\mathbf{r},\omega) = &\int G_0(\mathbf{r},\mathbf{r}_0)S(\mathbf{r}_0)\,dV_0 \\ &-\frac{\omega^2}{c_0^2}\iint G_0(\mathbf{r},\mathbf{r}_1)n(\mathbf{r}_1)G_0(\mathbf{r}_1,\mathbf{r}_0)S(\mathbf{r}_0)\,dV_1 dV_0 \\ &+\frac{\omega^4}{c_0^4}\iiint G_0(\mathbf{r},\mathbf{r}_2)n(\mathbf{r}_2)G_0(\mathbf{r}_2,\mathbf{r}_1)n(\mathbf{r}_1)G_0(\mathbf{r}_1,\mathbf{r}_0)S(\mathbf{r}_0)\,dV_2 dV_1 dV_0 \\ &+\cdots. \end{aligned} \tag{23.28}$$

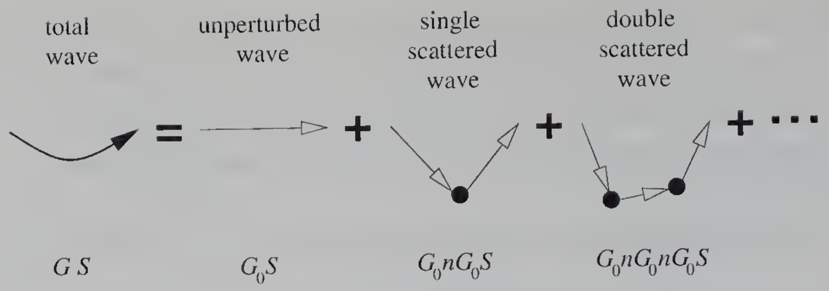

Fig. 23.2 Decomposition of the total wave field (thick solid line) in the unperturbed wave, the single scattered wave, the double scattered wave, and higher order scattering events. The total Green's function G is shown by a thick line, the unperturbed Green's function G_0 by thin lines, and each scattering event by the heterogeneity n is indicated by a solid dot.

This expansion is shown graphically in Figure 23.2. Reading each of the lines in expression (23.28) from right to left one can follow the "life-history" of the waves that are scattered in the medium. The top line of the right-hand side gives the unperturbed wave. This wave is excited by the source at location $\mathbf{r}_0$; this is described by the source term $S(\mathbf{r}_0)$. The wave then propagates through the unperturbed medium to the point $\mathbf{r}$; this is accounted for by the term $G_0(\mathbf{r}, \mathbf{r}_0)$. Graphically this is shown by the first diagram after the equality sign in Figure 23.2. In this figure the thin arrows denote the unperturbed Green's function G_0. The second line in (23.28) physically describes a wave that is generated at the source $S(\mathbf{r}_0)$; this wave then propagates through the unperturbed medium with the Green's function $G_0(\mathbf{r}_1, \mathbf{r}_0)$ to a scatterer at location $\mathbf{r}_1$. The wave is then scattered; this is accounted for by the terms $-\left(\omega^2/c_0^2\right) n(\mathbf{r}_1)$. In Figure 23.2 this scattering interaction is indicated by a solid dot. The wave then travels through the unperturbed medium to the point $\mathbf{r}$; this is accounted for by the term $G_0(\mathbf{r}, \mathbf{r}_1)$.

Problem f Describe in a similar way the "life-history" of the double scattered wave that is given by the last line of (23.28) and convince yourself that this corresponds to the right-most diagram in Figure 23.2.

The analysis in this section can be continued to any order. The resulting series is called the *Neumann series*. It gives a decomposition of the total wave field into single scattered waves, double scattered waves, triple scattered waves, etc. In practice it is often difficult to compute the waves that are scattered more than once. In the *Born approximation* one simply truncates the perturbation series after the second term. Using the notation of (23.25) this means that in the Born approximation the wave field is given by

$$p_B(\mathbf{r}) = p_0(\mathbf{r}) - \frac{\omega^2}{c_0^2} \int G_0(\mathbf{r}, \mathbf{r}') n(\mathbf{r}') p_0(\mathbf{r}')\, dV'. \qquad (23.29)$$

This expression is extremely useful for a large variety of applications. For sufficiently weak perturbations $\varepsilon n(\mathbf{r})$ it allows the analytical computation of the (single) scattered waves. In the Born approximation the scattered waves are given by the last term in (23.29). This last term gives a linear(ized) relation between the scattered waves and the perturbation of the medium. In many applications one measures the scattered waves and one wants to retrieve the perturbation of the medium. The Born approximation provides a linear relation between the scattered waves and the perturbation of the medium. Methods from linear algebra can then be used to infer the perturbation $n(\mathbf{r})$ of the medium from measurements of the scattered waves. The Born approximation provides the basis for most of the techniques used in reflection seismology for the detection of hydrocarbons in the Earth, see for example references [28] and [123]. The Born approximation also forms the basis of the imaging techniques used with radar [52] and a variety of other applications. In fact, it has been argued that imaging with multiple-scattered waves is not feasible in practice [28]; a discussion of this controversial issue can be found in reference [92].

In this section the Born approximation for the Helmholtz equation was derived. However, this derivation can readily be generalized to other scattering problems. The only required ingredient is that, when one divides the medium into an unperturbed medium and a perturbation, one can compute the Green's function for the unperturbed medium. The Born approximation is used in quantum mechanics [69], electromagnetic wave scattering [53], scattering of elastic body waves [122], and elastic surface waves [97].

There is a famous application of the Born approximation. According to (23.29) the scattered waves are multiplied by ω^2 compared to the unperturbed waves. This is also the case for the scattering of electromagnetic waves [53]. This term ω^2 explains why the sky is blue. The scattered waves are proportional to ω^2 compared to the unperturbed waves. This means that light with a high frequency is scattered more strongly than light with a lower frequency. In other words, blue light is scattered more strongly than red light. The blue light that comes from the Sun is scattered more effectively out of the light beam from the Sun to an observer than the red light. When this blue light is scattered again by small particles in the atmosphere it travels to an observer as blue light that comes from the sky from a different location than the Sun. We perceive this as "the blue sky." This argument should, however, be used with caution because the Green's function in (23.29) also depends on frequency.

23.3 Linear travel time tomography

An important tool for determining the interior structure of the Earth and other bodies is travel time tomography. In this technique one measures the travel time of waves between a large number of sources and receivers. When the coverage with rays is sufficiently dense, one can determine the velocity of seismic waves in the Earth from the recorded travel times. Detailed descriptions of seismic tomography can be found in references [51] and [75]. The travel time along a ray is given by the integral

$$\tau = \int \frac{1}{c(\mathbf{r})} ds. \tag{23.30}$$

Since the integral is proportional to $1/c(\mathbf{r})$, it is convenient to use the *slowness* $u(\mathbf{r}) = 1/c(\mathbf{r})$ rather than the velocity. Using this quantity the travel time is given by

$$\tau = \int_{\mathbf{r}[u]} u(\mathbf{r})\, ds. \tag{23.31}$$

The last expression suggests a linear relation between the measured travel time τ and the unknown slowness $u(\mathbf{r})$. Such a linear relation is ideal for solving the inverse problem of the determination of the slowness because one can resort to techniques from linear algebra. However, integral (23.31) is taken over the ray that joins the source and the receiver. The rays are curves of stationary travel time and themselves depend on the slowness. This dependence effectively makes the relation between the slowness and the travel time nonlinear. In this section we perturb both the slowness and the travel time to derive a linearized relation between the travel time *perturbation* and the slowness *perturbation*. The travel time along rays follows from geometric ray theory as shown in Section 12.4. It is shown in that section that the travel time $\tau(\mathbf{r})$ from a given source to location $\mathbf{r}$ is given by the eikonal equation (12.25) which can be written as

$$|\nabla\tau(\mathbf{r})|^2 = u^2(\mathbf{r}). \tag{23.32}$$

Let us assume that we have a reasonable guess $u_0(\mathbf{r})$ for the slowness, and that we seek a small perturbation of the slowness; this perturbation is denoted as $\varepsilon u_1(\mathbf{r})$. The slowness can then be written as

$$u(\mathbf{r}) = u_0(\mathbf{r}) + \varepsilon u_1(\mathbf{r}). \tag{23.33}$$

Again, the parameter ε only serves to systematically set up the perturbation treatment. When the slowness is perturbed, the travel time changes as well and it can be expanded in a perturbation series of the parameter ε:

$$\tau = \tau_0 + \varepsilon\tau_1 + \varepsilon^2\tau_2 + \cdots. \tag{23.34}$$

In this section we seek the relation between the first order travel time perturbation τ_1 and the slowness perturbation u_1.

Problem a Insert (23.33) and (23.34) into the eikonal equation (23.32), use that $|\nabla\tau|^2 = (\nabla\tau \cdot \nabla\tau)$ and collect together the terms in $O(1)$ and $O(\varepsilon)$ to show that the first and zeroth order travel time perturbation are given by

$$|\nabla\tau_0(\mathbf{r})|^2 = u_0^2(\mathbf{r}), \tag{23.35}$$

$$(\nabla\tau_0 \cdot \nabla\tau_1) = u_0 u_1. \tag{23.36}$$

The first equation is nothing but the eikonal equation for the unperturbed problem. This expression states that the length of the vector $\nabla\tau_0$ is equal to u_0. This means that $\nabla\tau_0$ can be written as

$$\nabla\tau_0 = u_0\hat{\mathbf{n}}_0. \tag{23.37}$$

In this expression the unit vector $\hat{\mathbf{n}}_0$ gives the direction of $\nabla\tau_0$ as shown in Figure 23.3. It was shown in Section 5.1 that the gradient $\nabla\tau_0$ is perpendicular to the surfaces of constant travel time, see Figure 23.3. It is also shown in that figure that the rays are the curves that are everywhere perpendicular to the surfaces of constant travel time.

Problem b Use (5.21) to show that the derivative of τ_0 along the $\mathbf{r}_0$ of the reference medium is given by

$$\frac{d\tau_0}{ds_0} = u_0\,, \tag{23.38}$$

where s_0 denotes the arclength along the ray in the reference medium.

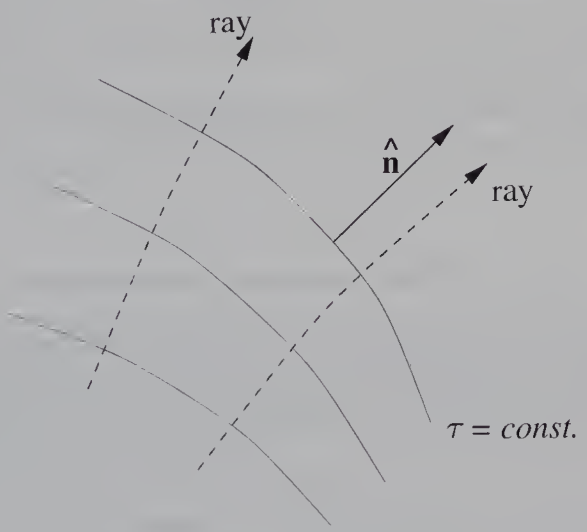

Fig. 23.3 Wavefronts as the surfaces of constant travel time τ (solid lines), and the rays that are the curves perpendicular to the travel time surfaces (dashed lines). The unit vector **n** is perpendicular to the wavefronts.

This last expression can be integrated to give

$$\tau_0 = \int_{\mathbf{r}_0[u_0]} u_0(\mathbf{r})\, ds_0. \tag{23.39}$$

This expression is identical to (23.31) with the exception that all quantities are for the reference medium u_0 and its associated rays $\mathbf{r}_0[u_0]$.

Problem c In order to derive the first order travel time perturbation, insert (23.37) into (23.36) to derive that

$$(\hat{\mathbf{n}}_0 \cdot \nabla\tau_1) = u_1. \tag{23.40}$$

Note that the unit vector $\hat{\mathbf{n}}_0$ is directed along the rays in the reference medium u_0 and that it is therefore independent of the slowness perturbation u_1.

Problem d Integrate the last expression to give

$$\tau_1 = \int_{\mathbf{r}_0[u_0]} u_1(\mathbf{r})\, ds_0. \tag{23.41}$$

In this expression the integration is along the rays $\mathbf{r}_0[u_0]$ in the reference medium. Since these rays are assumed to be known, (23.41) constitutes a linearized relation between the travel time perturbation τ_1 and the slowness perturbation u_1. Techniques from linear algebra can then be used to determine the unknown slowness perturbation from the measured travel time perturbations τ_1. In many textbooks (e.g. reference [75]) this result is derived from Fermat's theorem. However, the treatment in this section (which was proposed by Aldridge [4]) is conceptually much simpler. In fact, the treatment in this section can easily be extended to compute the travel time perturbation to *any* order [99].

23.4 Limits on perturbation theory

Perturbation theory is a powerful tool; in principle it provides a systematic way to derive the perturbation to any desired order. In this section we will discover that for a given order of truncation of the perturbation series the accuracy of the obtained result may depend strongly on the value of certain parameters of the problem that one is considering. This is illustrated with a simple problem which we can also solve analytically. We consider the differential equation

$$\ddot{x} + \omega_0^2(1+\varepsilon)x = 0, \tag{23.42}$$

with the initial conditions

$$x(0) = 1, \qquad \dot{x}(0) = 0. \tag{23.43}$$

This equation describes a harmonic oscillator in which the frequency is perturbed.

Problem a Show that the exact solution of this problem is given by

$$x(t) = \cos\left(\omega_0\sqrt{1+\varepsilon}\, t\right). \tag{23.44}$$

Problem b The solution $x(t)$ is a function of the perturbation parameter ε, it can therefore be written as a perturbation series in this parameter:

$$x(t) = x_0(t) + \varepsilon x_1(t) + \varepsilon^2 x_2(t) + \cdots. \tag{23.45}$$

Insert this series into (23.42) and collect together the terms of equal powers in ε to show that the terms $x_n(t)$ satisfy the following equations:

$$\left.\begin{aligned} \ddot{x}_0 + \omega_0^2 x_0 &= 0, \\ \ddot{x}_1 + \omega_0^2 x_1 &= -\omega_0^2 x_0, \\ \ddot{x}_2 + \omega_0^2 x_2 &= -\omega_0^2 x_1, \\ &\vdots \end{aligned}\right\} \tag{23.46}$$

Problem c In order to solve these equations one must also consider the initial conditions of $x_n(t)$. Obtain these condition by inserting the perturbation series (23.45) into the initial conditions (23.43) and derive that

$$\left.\begin{aligned} x_0(0) &= 1, & \dot{x}_0(0) &= 0, \\ x_n(0) &= 0, & \dot{x}_n(0) &= 0 \quad \text{for} \quad n \geq 1. \end{aligned}\right\} \tag{23.47}$$

Problem d Solve the differential equation for x_0 for the boundary condition of (23.47) and show that the solution is given by

$$x_0(t) = \cos(\omega_0 t). \tag{23.48}$$

Note that this expression is identical to the exact solution (23.44) when one switches off the perturbation by setting $\varepsilon = 0$. Inserting this solution into the second line of (23.46) one finds that the first order perturbation satisfies the following differential equation

$$\ddot{x}_1 + \omega_0^2 x_1 = -\omega_0^2 \cos(\omega_0 t). \tag{23.49}$$

This equation describes a harmonic oscillator with eigenfrequency ω_0 which is driven by a force on the right-hand side. This driving force also oscillates with frequency ω_0. This means that the oscillator x_1 is driven at its resonance frequency, which in general leads to a motion that grows with time.

Problem e In order to solve (23.49) write the perturbation $x_1(t)$ as

$$x_1(t) = f(t)\cos(\omega_0 t) + g(t)\sin(\omega_0 t). \tag{23.50}$$

Insert this expression into (23.49) and collect together the terms that multiply $\cos(\omega_0 t)$ and $\sin(\omega_0 t)$ to show that the unknown functions $f(t)$ and $g(t)$ obey the following differential equations:

$$\left.\begin{aligned} \ddot{f} + 2\omega_0\dot{g} &= -\omega_0^2, \\ \ddot{g} - 2\omega_0\dot{f} &= 0. \end{aligned}\right\} \tag{23.51}$$

Problem f Show that these equations are satisfied by the following solution:

$$\dot{f} = 0, \qquad \dot{g} = -\frac{\omega_0}{2}, \tag{23.52}$$

and integrate these equations to derive a particular solution that is given by

$$f = A, \qquad g(t) = -\frac{1}{2}\omega_0 t + B, \tag{23.53}$$

with A and B integration constants.

These expressions for f and g lead to the general solution

$$x_1(t) = -\frac{1}{2}\omega_0 t\ \sin(\omega_0 t) + A\cos(\omega_0 t) + B\sin(\omega_0 t). \tag{23.54}$$

It follows from (23.47) that the initial conditions for $x_1(t)$ are given by $x_1(0) = 0$, $\dot{x}(0) = 0$.

Problem g Derive from these initial conditions that the integration constants are given by $A = B = 0$, so that the solution is given by

$$x_1(t) = -\frac{1}{2}\omega_0 t\ \sin(\omega_0 t). \tag{23.55}$$

It was noted earlier that (23.49) describes an oscillator that is driven at its resonance frequency. Solution (23.55) grows linearly with time. This growth with time is called the *secular growth*. It is an artifact of the perturbation technique employed because the original problem (23.42) does not contain a resonant driving force at all. Inserting the first order perturbation (23.55) and the unperturbed solution (23.48) into the perturbation series (23.45) finally gives

$$x(t) = \cos(\omega_0 t) - \frac{\varepsilon}{2}\omega_0 t\ \sin(\omega_0 t) + O(\varepsilon^2). \tag{23.56}$$

Let us first verify that this expression is indeed the first order expansion of the exact solution (23.44). Using the series (3.12), (3.13), and (3.16), the exact solution (23.44) is to first order in ε given by

$$\begin{aligned} x(t) &= \cos\left(\omega_0\sqrt{1+\varepsilon}\ t\right) \\ &= \cos\left\{\omega_0\left[1+\frac{\varepsilon}{2}+O(\varepsilon^2)\right]t\right\} \\ &= \cos(\omega_0 t)\cos\left(\frac{\varepsilon}{2}\omega_0 t\right) - \sin(\omega_0 t)\sin\left(\frac{\varepsilon}{2}\omega_0 t\right) + O(\varepsilon^2) \\ &= \cos(\omega_0 t) - \frac{\varepsilon}{2}\omega_0 t\sin(\omega_0 t) + O(\varepsilon^2) \end{aligned} \tag{23.57}$$

This result is indeed identical to the first order expansion (23.56) that was obtained from perturbation theory. The fact that this result is correct, however, does not imply that this result is also useful. Perturbation theory is based on the premise that the truncated perturbation series gives a good approximation to the true solution. The truncation of this series only makes sense when the subsequent terms in the perturbation expansion rapidly become smaller. However, the first order term $(\varepsilon\omega_0 t/2)\sin(\omega_0 t)$ in (23.46) is as large as the zeroth order term $\cos(\omega_0 t)$ when $\varepsilon\omega_0 t/2 \sim 1$. This means that the first order perturbation series (23.56) ceases to be a good approximation to the true solution when

$$t \sim \frac{1}{\varepsilon\omega_0}. \tag{23.58}$$

The upshot of this example is that even though a truncated perturbation series may be correct, it may only be useful for a restricted range of parameters. In this example the first order perturbation series is a good approximation only when $t \ll 1/\varepsilon\omega_0$. For the problem in this section it would be more appropriate to make a perturbation series of the phase and the amplitude of the oscillator. Systematic techniques such as *multiple-scale analysis* [14] have been developed for this purpose. In the following section we carry out the transformation $p = e^S$ to derive the perturbations of the amplitude and phase of a wave that propagates through an inhomogeneous medium.

23.5 WKB approximation

In this section we analyze the propagation of a wave through a one-dimensional acoustic medium. In the frequency domain the pressure satisfies the following differential equation:

$$\rho\frac{d}{dx}\left(\frac{1}{\rho}\frac{dp}{dx}\right) + \frac{\omega^2}{c^2}p = 0. \tag{23.59}$$

In this expression both the density ρ and the velocity c vary with the position x. This equation can only be solved in closed form for a special form of the functions

$\rho(x)$ and $c(x)$. It should be noted that the treatment in this section is also applicable to the Schrödinger equation

$$\frac{d^2\psi}{dx^2} + \frac{2m}{\hbar^2}\left[E - V(x)\right]\psi = 0, \tag{23.60}$$

by making the substitutions

$$\rho(x) \rightarrow 1, \qquad p \rightarrow \psi, \qquad 1/c(x) \rightarrow \sqrt{2m\left(E - V(x)\right)}/\hbar\omega. \tag{23.61}$$

At this point there is as yet no small perturbation parameter. We restrict our attention to media in which the length scale of each variation in ρ and c is much larger than the wavelength λ of the wave. Physically this type of medium does not contain strong inhomogeneities on the scale of a wavelength, so that the waves are not reflected strongly by the heterogeneity. Let the length scale of the heterogeneity be denoted by L. When this length scale is much larger than the wavelength $\lambda = 2\pi\omega/c$, the following parameter is small:

$$\varepsilon = \frac{c}{\omega L} \ll 1. \tag{23.62}$$

As noted in the previous section, when the perturbation affects mostly the phase of a wave, it is advantageous to perturb the phase (and amplitude) of the wave rather than the solution p itself. This can be achieved by making the transformation

$$p = e^S. \tag{23.63}$$

When S is complex this transformation is without any loss of generality.

Problem a Insert this relation into (23.59) and derive that $S(x)$ satisfies the following differential equation:

$$\frac{d^2S}{dx^2} - \frac{1}{\rho}\frac{d\rho}{dx}\frac{dS}{dx} + \left(\frac{dS}{dx}\right)^2 + \frac{\omega^2}{c^2} = 0. \tag{23.64}$$

At this point we have made the problem more complex because this equation is nonlinear in the unknown function $S(x)$ whereas the original equation (23.59) is linear in the pressure $p(x)$. However, we have not yet applied the perturbation technique. Before we do this, let us first reflect on the transformation (23.63). If the medium were homogeneous, the solution would be given by $p = Ae^{ikx}$, with the wavenumber given by $k = \omega/c$. This special solution corresponds to $S = \ln(A) + ikx$, so that $dS/dx = ik = i\omega/c$. For an inhomogeneous medium one may expect the derivative of the phase to be close to this value, therefore we make the following substitution:

$$\frac{dS}{dx} = \frac{i\omega}{c(x)} F(x). \tag{23.65}$$

Problem b Show that this transformation transforms (23.64) into the following differential equation for F:

$$\frac{dF}{dx} - \frac{1}{\rho c}\frac{d(\rho c)}{dx} F + \frac{i\omega}{c} F^2 = \frac{i\omega}{c}. \tag{23.66}$$

Now we use perturbation analysis by using that the parameter ε defined in (23.62) is much smaller than unity. This can be achieved by transforming the distance x to a dimensionless distance ξ defined by

$$\xi \equiv x/L, \tag{23.67}$$

where L is the characteristic length scale of the velocity and density variations.

Problem c Under this transformation the derivative d/dx changes to $d/dx = d/d(\xi L) = (1/L)\, d/d\xi$. Use this to show that $F(\xi)$ satisfies the following differential equation:

$$\frac{c}{\omega L}\frac{dF}{d\xi} - \frac{c}{\omega L}\frac{1}{\rho c}\frac{d(\rho c)}{d\xi} F + iF^2 = i. \tag{23.68}$$

This equation contains the small dimensionless parameter $c/\omega L$ defined in (23.62), so that this equation is equivalent to

$$\varepsilon\frac{dF}{d\xi} - \varepsilon\frac{1}{\rho c}\frac{d(\rho c)}{d\xi} F + iF^2 = i. \tag{23.69}$$

Now we have an equation that looks similar to those in the perturbation problems we have seen in this chapter. We solve this equation by inserting the following perturbation series for $F(\xi)$:

$$F(\xi) = F_0(\xi) + \varepsilon F_1(\xi) + \cdots. \tag{23.70}$$

Problem d Insert this perturbation series into the differential equation (23.69) to derive that $F_0(\xi)$ and $F_1(\xi)$ satisfy the following equations:

$$\left.\begin{aligned} F_0^2 &= 1, \\ 2iF_0F_1 &= -\frac{dF_0}{d\xi} + \frac{1}{\rho c}\frac{d(\rho c)}{d\xi} F_0. \end{aligned}\right\} \tag{23.71}$$

The first of these equations has the solutions $F_0 = \pm 1$. According to (23.65) this corresponds to the phase derivative $dS/dx = \pm i\omega/c$. The plus sign denotes a right-going wave, and the minus sign a left-going wave. We focus here on a right-going wave so that $F_0 = +1$.

Problem e Insert this solution into the second line of (23.71) and show that $F_1(x')$ is given by

$$F_1(\xi) = -\frac{i}{2}\frac{1}{\rho c}\frac{d(\rho c)}{d\xi}. \tag{23.72}$$

This means that the first-order perturbation series for $F(\xi)$ is given by

$$F(\xi) = 1 - \frac{i\varepsilon}{2}\frac{1}{\rho c}\frac{d(\rho c)}{d\xi} + O(\varepsilon^2). \tag{23.73}$$

Problem f Now that we have obtained this solution as a function of the transformed distance ξ, transform back to the original distance x by using (23.67). Use (23.62) to show that the solution (23.73) is equivalent to

$$F(x) = 1 - \frac{i}{2}\frac{1}{\rho\omega}\frac{d(\rho c)}{dx} + \cdots. \tag{23.74}$$

Problem g Use (23.65) to convert this into an equation for dS/dx. Integrate this equation to show that the solution $S(x)$ is given by

$$S(x) = i\int_{-\infty}^{x}\frac{\omega}{c(x')}dx' + \frac{1}{2}\ln\left[\rho(x)c(x)\right] + B. \tag{23.75}$$

Problem h Use the transformation (23.63) to show that this solution corresponds to the following pressure field

$$p(x) = A\sqrt{\rho(x)c(x)}\,\exp\left(i\int_{-\infty}^{x}\frac{\omega}{c(x')}dx'\right), \tag{23.76}$$

with the new constant defined by $A = e^B$.

This solution states that the wave propagates to the right with an amplitude that is proportional to $\sqrt{\rho(x)c(x)}$. The local wavenumber $k(x)$ of the wave is given by the derivative of the phase of the wave [119]; it is therefore given by

$$k(x) = \frac{d}{dx}\int_{-\infty}^{x}\frac{\omega}{c(x')}dx' = \frac{\omega}{c(x)}. \tag{23.77}$$

This means that the local wavenumber at a location x is given by the wavenumber $\omega/c(x)$ that the medium would have if it were homogeneous with the properties of the medium at that location x. The solution (23.76) is known as the WKB solution (named after Wentzel, Kramers, and Brillouin). Seismologists prefer to call this solution the WKBJ approximation because of the contribution of Lord Jeffreys [54].

In most textbooks (e.g. reference [69]) this solution is derived for the Schrödinger equation rather than the acoustic wave equation; with the transformation (23.61) the derivations are equivalent.

Problem i Use the correspondence (23.61) to show that in quantum mechanics the WKB solution is given by

$$\psi(x) = \frac{A}{[E - V(x)]^{1/4}} \exp\left\{ i \int_{-\infty}^{x} \frac{\sqrt{2m\,[E - V(x')]}}{\hbar} dx' \right\}. \tag{23.78}$$

Problem j Show that this approximation is infinite at the *turning points* of the motion. These are the points where the total energy of the particle is equal to the potential energy. This means that the WKB solution breaks down at the turning points.

A clear account of the WKB approximation with a large number of applications is given by Bender and Orszag [14].

23.6 Need for consistency

In perturbation theory one derives an approximate solution to a problem. In many applications this approximation is then used in subsequent calculations. In doing so, one must keep in mind that the solution obtained from perturbation theory is not the true solution, and that it is pointless to carry out the subsequent calculations with an accuracy which is higher than the accuracy of the solution obtained from perturbation theory.

As an example we consider in this section the WKB solution (23.76) for the pressure field $p(x)$ and compute the particle velocity $v(x)$ that is associated with this pressure field. We assume that the motion is sufficiently small that the equation of motion can be linearized so that Newton's law gives $\rho \partial v/\partial t = F$. As shown in Section 5.2 the pressure force is given by $F = -\partial p/\partial x$, so that in the time domain Newton's law is given by $\rho \partial v/\partial t = -\partial p/\partial x$.

Problem a Show that with the Fourier convention (15.42) the corresponding equation is given in the frequency domain by

$$i\omega\rho v = \partial p/\partial x. \tag{23.79}$$

Problem b Apply this result to the WKB solution (23.76) and show that the velocity is given by

$$i\omega\rho v = \frac{1}{2}\frac{1}{\rho c}\frac{d(\rho c)}{dx}p + \frac{i\omega}{c}p. \tag{23.80}$$

Problem c Use the estimate of the derivative in Section 12.2 to show that the first term on the right-hand side is of the order p/L, where L is the characteristic length scale over which the density and the velocity vary.

The second term on the right-hand side of (23.80) is of the order $\omega p/c$. This means that the ratio of the first term to the second term is given by $(p/L)/(\omega p/c) = c/(\omega L) = \varepsilon$, where the parameter ε is defined in (23.62). In the previous section we assumed that the medium varies so smoothly that $\varepsilon \ll 1$. This means that under the assumptions which underlie the WKB approximation the first term on the right-hand side of (23.80) can be ignored with respect to the second term.

Problem d Show that in this approximation the velocity is given by

$$v = \frac{p}{\rho c}. \tag{23.81}$$

Note that ignoring the first term is consistent with the terms that we have ignored in the WKB approximation.

This last expression has an interesting interpretation. The quantity ρc is called the *acoustic impedance*. This term is reminiscent of the theory of electromagnetism. For a resistor, Ohm's law $I = V/R$ relates the current I that is generated by a voltage V. For a general linear electric system the current and the voltage are related by

$$I = \frac{V}{Z}, \tag{23.82}$$

where Z is a generalization of the resistance that is called the *impedance*. The impedance gives the strength of the current for a given potential. Similarly, the acoustic impedance ρc in (23.81) gives the particle velocity for a given pressure.

Combining (23.76) with (23.81) shows that the velocity, and hence the particle motion, is proportional to $1/\sqrt{\rho c}$. This means that the particle motion increases when the acoustic impedance decreases. This has important implications for earthquake hazards. For soft soils, both the density ρ and the wave velocity c are small. This means that the acoustic impedance is much smaller in soft soils than in hard rock. This in turn means that the ground motion during earthquakes is much more severe in soft soils than in hard rock. (The motion in the Earth is governed by the elastic wave equation rather than the acoustic wave equation. However, one can show [3] that for elastic waves also the displacement is inversely proportional to $1/\sqrt{\rho c}$, where c is the propagation velocity of the elastic wave under consideration.)

The fact that the ground motion is inversely proportional to the square-root of the impedance is one of the factors that made the 1985 earthquake along the west coast of Mexico cause so much damage in Mexico City. This city is constructed on soft sediments which have filled the swamp onto which the city is built. The small value of the associated elastic impedance was one of the causes of the extensive damage in Mexico City after this earthquake.

23.7 Singular perturbation theory

In Section 23.1 we analyzed the behavior of the root of the equation $x^3 - 4x^2 + 4x = \varepsilon$ that was located near $x = 0$. As shown in that section, the unperturbed problem also has a root $x = 2$. The roots $x = 0$ and $x = 2$ can be seen graphically in Figure 23.1 because for these values of x the polynomial shown by the thick solid line is equal to zero. In Figure 23.1 the value $\varepsilon = +0.15$ is shown by a dotted line while the value $\varepsilon = -0.15$ is indicated by the dashed line. There is a profound difference between the two roots when the parameter ε is nonzero. The root near $x = 0$ depends in a continuous way on ε, and (23.2) has for the root near $x = 0$ a solution regardless of whether ε is positive or negative. This situation is completely different for the root near $x = 2$. When ε is positive (the dotted line), the polynomial has *two* intersections with the dotted line, whereas when ε is negative the polynomial does *not* intersect the dashed line at all. This means that depending on whether ε is positive or negative, the solution has two or zero solutions, respectively. This behavior cannot be described by a regular perturbation series of the form (23.5) because this expansion assigns *one* solution to each value of the perturbation parameter ε.

Let us first diagnose where the treatment of Section 23.1 breaks down when we apply it to the root near $x = 2$.

Problem a Insert the unperturbed solution $x_0 = 2$ into the second line of (23.13) and show that the resulting equation for x_1 is

$$0 \cdot x_1 = 1. \tag{23.83}$$

This equation obviously has no finite solution. This is related to the fact that the tangent of the polynomial at $x = 2$ is horizontal. First-order perturbation theory effectively replaces the polynomial by the straight line that is tangent to the polynomial. When this tangent line is horizontal, it can never have a value that is nonzero.

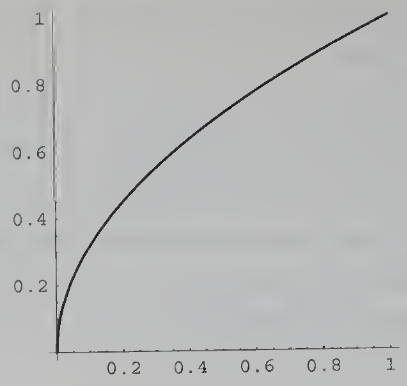

Fig. 23.4 Graph of the function $\sqrt{\varepsilon}$.

This means that the regular perturbation series (23.5) is not the appropriate way to study the behavior of the root near $x = 2$. In order to find out how this root behaves, let us set

$$x = 2 + y. \tag{23.84}$$

Problem b Show that under the substitution (23.84) the original problem (23.2) transforms to

$$y^3 + 2y^2 = \varepsilon. \tag{23.85}$$

We will not yet carry out a systematic perturbation analysis, but we will first determine the dependence of the solution y on the parameter ε. For small values of ε, the parameter y is also small. This means that the term y^3 can be ignored with respect to the term y^2. Under this assumption (23.85) is approximately equal to $2y^2 \approx \varepsilon$ so that $y \approx \sqrt{\varepsilon/2}$. This means that the solution does not depend on integer powers of ε as in the perturbation series (23.5), but that it does depend on the square-root of ε. The square-root of ε is shown in Figure 23.4. Note that for $\varepsilon = 0$ the tangent of this curve is vertical and that for $\varepsilon < 0$ the function $\sqrt{\varepsilon}$ is not defined for real values of ε .[‡] This reflects the fact that the roots near $x = 2$ depend in a very different way on ε than the root near $x = 0$.

We know now that a regular perturbation series (23.5) is not the correct tool to use to analyze the root near $x = 2$. However, we do not yet know what type of perturbation series we should use for the root near $x = 2$; we only know that the perturbation depends to leading order on $\sqrt{\varepsilon}$. That is, let us make the following substitution:

[‡] When one allows a complex solution $x(\varepsilon)$ of the equation, there are always two roots near $x = 2$. However, these complex solutions also display a fundamental change in their behavior when $\varepsilon = 0$, which is characterized by a bifurcation.

$$x = 2 + \sqrt{\varepsilon}\, z. \tag{23.86}$$

Problem c Insert this solution into (23.2) and show that z satisfies the following equation:

$$\sqrt{\varepsilon} z^3 + 2z^2 = 1. \tag{23.87}$$

Now we have a new perturbation problem with a small parameter. However, this small parameter is not the original perturbation parameter ε, but it is the square-root $\sqrt{\varepsilon}$. The perturbation problem in this section is a *singular perturbation problem.* In a singular perturbation problem the solution is not a well-behaved function of the perturbation parameter. This has the result that the corresponding perturbation series cannot be expressed in powers ε^n, where n is a positive real integer. Instead, negative or fractional powers of ε are present in the perturbation series of a singular perturbation problem.

Problem d Since the small parameter in (23.87) is $\sqrt{\varepsilon}$, it makes sense to seek an expansion of z in this parameter:

$$z = z_0 + \varepsilon^{1/2} z_1 + \varepsilon z_2 + \cdots. \tag{23.88}$$

Collect together the coefficients of equal powers of ε when this series is inserted into (23.87) and show that this leads to the following equations for the coefficients z_0 and z_1:

$$\left.\begin{array}{ll} O(1)\text{-terms:} & 2z_0^2 - 1 = 0, \\ O(\varepsilon^{1/2})\text{-terms:} & z_0^3 + 4z_0 z_1 = 0. \end{array}\right\} \tag{23.89}$$

Problem e The first equation of (23.89) obviously has the solution $z_0 = \pm 1/\sqrt{2}$. Show that for both the plus and the minus signs $z_1 = -1/2$. Use these results to derive that the roots near $x = 2$ are given by

$$x = 2 \pm \frac{1}{\sqrt{2}}\sqrt{\varepsilon} - \frac{1}{2}\varepsilon + O(\varepsilon^{3/2}). \tag{23.90}$$

It is illustrative to compute the numerical values of these roots for the original problem (23.1), where $\varepsilon = 0.01$; this gives for the two roots:

$$x = 1.924 \qquad \text{and} \qquad x = 2.065. \tag{23.91}$$

In these numbers only three decimals are shown. The reason is that the error in the truncated perturbation series is of the order of the first truncated term, hence the error is of the order $(0.01)^{3/2} = 0.001$. When these solutions are compared with the perturbation solution (23.16) for the root near $x = 0$, it is striking that the singular perturbation series for the root near $x = 2$ converges much less rapidly than the regular perturbation series (23.16) for the root near $x = 0$. This is a consequence of the fact that the solution near $x = 2$ is a perturbation series in $\sqrt{\varepsilon}(= 0.1)$ rather than $\varepsilon(= 0.01)$. When the roots (23.91) are inserted into the polynomial (23.1) the following solutions are obtained for the two roots:

$$\left.\begin{array}{ll} x = 1.924: & x^3 - 4x^2 + 4x = 0.0111 = 0.01 + 0.0011, \\ x = 2.065: & x^3 - 4x^2 + 4x = 0.0087 = 0.01 - 0.0012. \end{array}\right\} \qquad (23.92)$$

Note that these results are much less accurate than the corresponding result (23.17) for the root near $x = 0$. Again this is a consequence of the singular behavior of the roots near $x = 2$.

The singular behavior of the roots of the polynomial (23.1) near $x = 2$ corresponds to the fact that the solution changes in a discontinuous way when the perturbation parameter ε goes to zero. It follows from Figure 23.1 that for the perturbation problem in this section the problem has one root near $x = 2$ when $\varepsilon = 0$, it has no roots when $\varepsilon < 0$ and there are two roots when $\varepsilon > 0$. Such a discontinuous change in the character of the solution also occurs in fluid mechanics in which the equation of motion is given by

$$\frac{\partial(\rho\mathbf{v})}{\partial t} + \nabla\cdot(\rho\mathbf{v}\mathbf{v}) = \mu\nabla^2\mathbf{v} + \mathbf{F}. \qquad (11.55)$$

In this expression the viscosity of the fluid gives a contribution $\mu\nabla^2\mathbf{v}$, where μ is the viscosity. This viscous term contains the highest spatial derivatives of the velocity that are present in the equation. When the viscosity μ goes to zero, the equation for fluid flow becomes a first order differential equation rather than a second order differential equation. This changes the number of boundary conditions that are needed for the solution, and hence it drastically affects the mathematical structure of the solution. This has the effect that boundary-layer problems are, in general, singular perturbation problems [111].

When waves propagate through an inhomogeneous medium they may be focused onto focal points or focal surfaces [16]. These regions in space where the wave amplitude is large are called *caustics*. The formation of caustics depends on $\varepsilon^{2/3}$, where ε is a measure of the variations in the wave velocity [58, 102]. The non-integer power of ε indicates that the formation of caustics constitutes a singular perturbation problem.

24

Asymptotic evaluation of integrals

In mathematical physics, the result of a computation often is expressed as an integral. Frequently, an analytical solution to these integrals is not known or it is so complex that it gives little insight into the physical nature of the solution. As an alternative, one can often find an approximate solution to the integral that in many cases works surprisingly well. The approximations that are treated here exploit that there is a parameter in the problem that is either large or small. The corresponding approximation is called an asymptotic solution to the integral, because the approximation holds for asymptotically large or small values of that parameter. Excellent treatments of the asymptotic evaluation of integrals are given by Bender and Orszag [14] and by Bleistein and Handelsman [18].

24.1 Simplest tricks

In general there is no simple recipe for integrating a function. For this reason, there is no simple trick for approximating integrals that always works. The asymptotic evaluation of integrals is more like a bag of tricks. In this section we treat the simplest tricks; the Taylor series and integration by parts. As an example, let us consider the following integral

$$F(x) \equiv \int_0^x e^{-u^2} du. \tag{24.1}$$

For small values of x, the integrand can be approximated well by a Taylor series. Each term in this Taylor series can be integrated separately.

Problem a Expand the integrand in a Taylor series around $u = 0$, and integrate this series term by term to derive that

$$\int_0^x e^{-u^2} du = x - \frac{1}{3}x^3 + \frac{1}{10}x^5 - \frac{1}{42}x^7 + \cdots \tag{24.2}$$

Problem b Show that this result can also be written as

$$\int_0^x e^{-u^2} du = \sum_{n=0}^{\infty} \frac{(-1)^n}{(2n+1)\,n!} x^{2n+1}. \tag{24.3}$$

In general, such an infinite series is just as difficult to evaluate as the original integral. The approximation to this integral consists in truncating the integral at a certain point. For example, retaining the first two terms gives the approximation

$$\int_0^x e^{-u^2} du \approx x - \frac{1}{3} x^3. \tag{24.4}$$

The approximation sign $\approx$ is mathematically not very precise, for this reason the following notation is preferred by many

$$\int_0^x e^{-u^2} du = x - \frac{1}{3} x^3 + O(x^5). \tag{24.5}$$

This notation has the advantage that is shows that the error made in (24.4) goes to zero as x^5 (or faster) as $x \to 0$. This type of mathematical rigor may appear to be attractive. However, the notation (24.5) still does not tell us how good the approximation (24.4) is for a given *finite* value of x. Since the mathematical rigor of (24.5) is not very informative for this problem, we will often use the more sloppy, but equally uninformative, notation (24.4). We return to this issue at the end of this section when we apply this result to the computation of probabilities for the Gaussian distribution.

In the following example we consider the integral

$$I(x) \equiv \int_x^{\infty} e^{-u^2} du, \tag{24.6}$$

for $x \gg 1$. The approximation that we will derive is based on the following identity:

$$e^{-u^2} = \frac{-1}{2u} \frac{de^{-u^2}}{du}. \tag{24.7}$$

Problem c Show this identity.

Problem d Insert (24.7) into (24.6) and carry out an integration by parts to show that

$$I(x) = \frac{1}{2x} e^{-x^2} - \frac{1}{2} \int_x^{\infty} \frac{1}{u^2} e^{-u^2} du. \tag{24.8}$$

For $x \gg 1$, the integration variable u is also much larger than one, and $1/u^2 \ll 1$. For this reason, the integral in (24.8) is much smaller than the original integral (24.6).

The first term in the right-hand side of (24.8) therefore is a good approximation to the integral:

$$\int_x^\infty e^{-u^2}\,du \approx \frac{1}{2x}e^{-x^2}. \tag{24.9}$$

However, this approximation can be refined further.

Problem e Insert the identity (24.7) into the integral in (24.8), carry out another integration by parts to show that

$$I(x) = \frac{1}{2x}e^{-x^2} - \frac{1}{4x^3}e^{-x^2} - \frac{1}{12}\int_x^\infty \frac{1}{u^4}e^{-u^2}\,du. \tag{24.10}$$

Since $u \gg 1$, the integral in (24.10) is much smaller than the integral in (24.8). The result of problem e leads to the two-term approximation

$$\int_x^\infty e^{-u^2}\,du = \frac{1}{2x}e^{-x^2}\left(1 - \frac{1}{2x^2} + O\left(\frac{1}{x^4}\right)\right). \tag{24.11}$$

In going from (24.10) to the previous expression we used that

$$\int_x^\infty \frac{1}{u^4}e^{-u^2}\,du \le \frac{1}{x^4}\int_x^\infty e^{-u^2}\,du \le \frac{1}{x^4}\int_0^\infty e^{-u^2}\,du = \frac{\sqrt{\pi}}{2x^4}. \tag{24.12}$$

The tricks shown in this section can be applied to many other integrals. Here we discuss an application of the approximations (24.5) and (24.11). Suppose that a random variable y follows a Gaussian distribution with zero mean and standard deviation σ. The corresponding probability density function is then given by

$$p(y) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{y^2}{2\sigma^2}\right). \tag{24.13}$$

The probability that y lies between a and b is given by

$$P(a < y < b) = \int_a^b p(y)\,dy. \tag{24.14}$$

The integral of (24.13) is not known in closed form. Tables of the integral of the Gaussian distribution exist. Bevington and Robinson [12] give a table for the probability $P(|y| < a\sigma)$. For example, the probability that $|y| < \sigma/2$ is equal to 0.352. We can estimate this probability with the small-x expansion (24.4) by using that

$$P(|y| < a\sigma) = 2P(0 < y < a\sigma) = (2/\sqrt{2\pi}\sigma)\int_0^{a\sigma} \exp(-y^2/2\sigma^2)\,dy. \tag{24.15}$$

Table 24.1 *The probability $P(|y| < \sigma/2)$ and various approximations to this probability.*

$P(\|y\| < \sigma/2)$		error (%)
True value	0.352	
Leading order asymptotic expansion	0.398	13
Second order asymptotic expansion	0.382	8.6

Table 24.2 *The probability $P(y > 3\sigma)$ and various approximations to this probability.*

$P(y > 3\sigma)$		error (%)
True value	0.001 35	
Leading order asymptotic expansion	0.001 48	9
Second order asymptotic expansion	0.001 32	2.5

Problem f Use this result to compute the probability $P(|y| < \sigma/2)$ with the one-term Taylor expansion and the two-term Taylor expansion (24.4) and show that these approximations give the estimated probabilities shown in Table 24.1.

Note that the Taylor approximation gives a reasonably accurate estimate for this probability. Taking more terms of the Taylor series into account leads to more accurate estimates of this probability.

The large-x expansion (24.11) can be used to estimate the probability $P(y > a\sigma)$ that y exceeds a given value $a\sigma$. Here we estimate the probability that y is larger than three standard deviations from the mean by using that $P(y > 3\sigma) = (1/\sqrt{2\pi}\sigma)\int_{3\sigma}^{\infty} \exp(-y^2/2\sigma^2)dy$. According the Bevington and Robinson [12] this probability is equal to 0.001 35.

Problem g Use the large-x expansion (24.11) to show that the leading order and second order asymptotic expansion give the estimates of this probability shown in Table 24.2.

Note that the estimates from the asymptotic expansion (24.11) are quite accurate, despite the fact that the employed value in its argument was $x = 3/\sqrt{2} = 2.12$. It is difficult to sustain that this value of x is much larger than one!

24.2 What does $n!$ have to do with e and $\sqrt{\pi}$?

The factorial function is defined as

$$n! \equiv 1 \cdot 2 \cdot 3 \cdots (n-1) \cdot n. \tag{24.16}$$

In this section we investigate how $n!$ behaves for large values of n. At first you might be puzzled that we seek an approximation to this function, because the recipe (24.16) is simple. However, the statistical mechanics of a many-particle system depends on the number of ways in which particles can be permuted [62]. Since a macroscopic system typically contains a mole of particles, one needs to evaluate $n!$ when n is of the order 10^{23}. For these types of problem the asymptotic behavior of $n!$ is very useful.

In order to derive this asymptotic behavior we first express $n!$ in an integral by using the gamma function that is defined as

$$\Gamma(x) \equiv \int_0^\infty u^{x-1} e^{-u} du. \tag{24.17}$$

With an integration by parts this function can be rewritten as

$$\Gamma(x+1) = -\int_0^\infty u^x \frac{de^{-u}}{du} du = -\left[u^x e^{-u}\right]_{u=0}^{u=\infty} + \int_0^\infty \frac{du^x}{du} e^{-u} du. \tag{24.18}$$

Problem a Evaluate the boundary term in the right-hand side and use the definition (24.17) to show that

$$\Gamma(x+1) = x\Gamma(x). \tag{24.19}$$

Problem b Show by direct integration that $\Gamma(1) = 1$. Use this result and (24.19) to show that when n is a positive integer

$$\Gamma(n+1) = n! \tag{24.20}$$

The combination of (24.17) and (24.20) implies that $n!$ can be expressed in the following integral

$$n! = \int_0^\infty F_n(u)\, du, \tag{24.21}$$

with

$$F_n(u) = u^n e^{-u}. \tag{24.22}$$

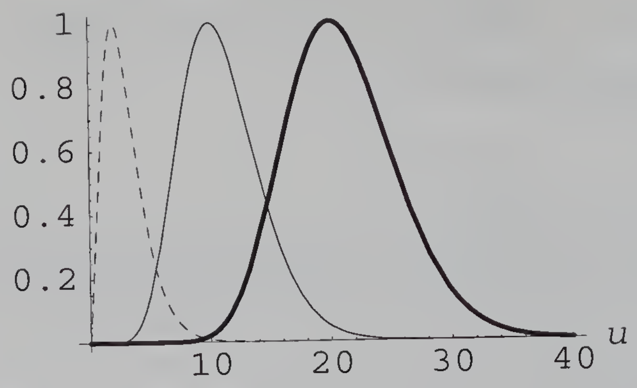

Fig. 24.1 The function $F_n(u)$ normalized by its maximum value for $n = 2$ (dashed line), $n = 10$ (thin solid line), and $n = 20$ (thick solid line).

It is instructive to consider the behavior of this function for large values of n. The term u^n is a rapidly growing function of u while the term e^{-u} is a decreasing function of u. For small values of u the first term dominates, while for large values of u the last term dominates. This means that $F_n(u)$ must have a maximum for some intermediate value of u.

Problem c Show that $F_n(u)$ has a maximum for $u_{\max} = n$.

The function $F_n(u)$ is shown in Figure 24.1 for $n = 2$, 10, and 20. The maximum of the function occurs for $u = n$. Note that as n increases, $F_n(u)$ becomes more and more a symmetrically peaked function around its maximum.

Since for large values of n the integrand $F_n(u)$ is a peaked function, the value of the integral will mostly be determined by the behavior of the integrand near the peak. Let us first consider a Taylor approximation of the function near its maximum.

Problem d Show that the second order Taylor approximation around its maximum $u = n$ is given by

$$F_n(u) = n^n e^{-n} - \frac{1}{2}(n-1)n^{n-1}e^{-n}(u-n)^2. \tag{24.23}$$

What we have done in this expression is approximate the function $F_n(u)$ by a parabola. As shown by the dashed line in Figure 24.2, for $n = 20$ this is good approximation of the function near its maximum, but this is not a good approximation of the tails of the function as it approaches zero. The second order Taylor approximation (24.23) does not account for the fact that $F_n(u)$ roughly has the shape of a Gaussian (which is defined as $\exp(-a(u-u_0)^2)$). In order to describe the Gaussian shape of the integrand we carry out a Taylor expansion of the exponent

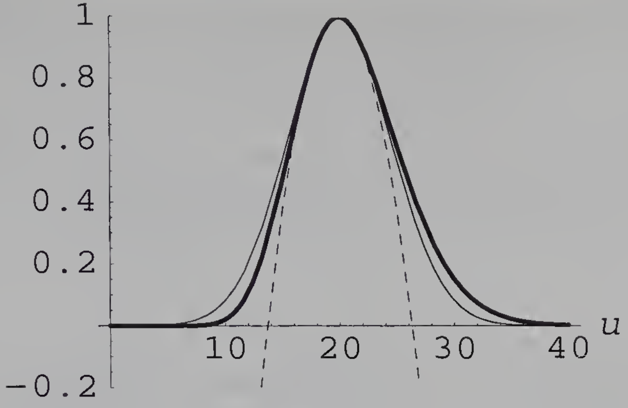

Fig. 24.2 The function $F_{20}(u)$ normalized by its maximum value (thick solid line), the parabolic approximation (dashed line), and the Gaussian approximation (thin solid line).

of the function rather than the original function by defining

$$h_n(u) = \ln\left(F_n(u)\right). \tag{24.24}$$

Problem e Show that

$$h_n(u) = -u + n \ln u , \tag{24.25}$$

and that

$$n! = \int_0^\infty e^{h_n(u)} du. \tag{24.26}$$

Problem f The maximum of $F_n(u)$ at $u = n$ is also the maximum of $h_n(u)$. Use a second order Taylor expansion of $h_n(u)$ around its maximum to derive that

$$n! \approx n^n e^{-n} \int_0^\infty \exp(-\frac{1}{2n}(u-n)^2)\, du. \tag{24.27}$$

The approximation sign means that terms $O((u-n)^3)$ in the exponent have been ignored.

Problem g The integrand of this integral is a Gaussian function with its peak at $u = n$. Show that the width of the peak (the "standard deviation") is given by $\sqrt{n}$.

Problem h The last result implies that the width of the Gaussian ($\sqrt{n}$) is for large values of n small compared to the location of the maximum (n). Compute the value of the Gaussian function at the lower end of the integration interval ($u = 0$) for $n = 5$, 10, and 100, respectively.

The integral (24.27) runs from $u = 0$ to $u = \infty$. If the integration ran from $u = -\infty$ to $u = \infty$ then we could solve the integral using

$$\int_{-\infty}^{\infty} e^{-ax^2} dx = \sqrt{\frac{\pi}{a}}. \tag{24.28}$$

However, as shown in Figure 24.2 the integrand in (24.27) is for large values of n small for negative values of u. Also, as shown in the last problem, even for moderately large values of n, the integrand is extremely small at the integration boundary $u = 0$. For this reason, we make a small error by extending the integration limit in (24.27) to $-\infty$:

$$n! \approx n^n e^{-n} \int_{-\infty}^{\infty} \exp(-\frac{1}{2n}(u-n)^2)\, du. \tag{24.29}$$

Problem i Use the Gaussian integral (24.28) to show that

$$n! \approx n^n e^{-n} \sqrt{2\pi n}. \tag{24.30}$$

This approximation is called *Stirling's formula*. In this expression the meaning of the $\approx$ sign is not precise. In the derivation of (24.30) we have made two approximations; the second order Taylor approximation of problem f and the extension of the lower limit of the integral (24.27) from 0 to $-\infty$. The first approximation can be improved by using a higher-order Taylor approximation, while the second approximation can be improved by using an asymptotic expansion of the error function as presented in the previous section. A more detail analysis of this problem shows that $n!$ can be written as [14]:

$$n! \approx n^n e^{-n} \sqrt{2\pi n} \left(1 + \frac{c_1}{n} + \frac{c_2}{n^2} + \cdots\right). \tag{24.31}$$

This series behaves differently to the series we have encountered so far since it has the following properties [14]:

- For small n the c_n decrease with increasing values of n. However, for larger values of n the c_n increase so rapidly with n that the series diverges. Trying to sum the series therefore is pointless.
- When the series is truncated at a certain point, the truncated series becomes more and more accurate as n increases.
- For a given value of n there is an optimal truncation point.

Table 24.3 *The values of $n!$ and its approximation by Stirling's formula for several values of n.*

n	$n!$	$n^n e^{-n} \sqrt{2\pi n}$	relative error (%)
1	1	0.922	7.8
2	2	1.919	4.1
3	6	5.836	2.7
4	24	23.51	2.1
5	120	118.0	1.6
10	3.629×10^6	3.598×10^6	0.83
15	1.307×10^{12}	1.300×10^{12}	0.55
20	2.432×10^{19}	2.423×10^{19}	0.42

With these remarks you may have some reservation about using the approximation (24.30). Table 24.3 shows this approximation to $n!$ for several values of n. For $n \geq 10$ the error made by (24.30) is less than 1%. The approximation (24.30) is made for the case $n \gg 1$. However, for $n = 1$, the approximation is accurate within 8%, whereas one certainly cannot maintain that in that case $n \gg 1$. Try to imagine how accurate the approximation is when n is equal to Avogadro's number (6×10^{23})! This is the number typically used in application of (24.30) in statistical mechanics. Lavenda [62] gives a wonderful historical account of how Max Planck discovered "Planck's law." A crucial step in his analysis was that he asked himself in how many ways a given amount of energy can be divided among a number of oscillators that each carry a given amount of energy. This question leads to the introduction of the binomial coefficients $\binom{n}{m} \equiv n!/m!(n-m)!$. Analyzing the factorials with Stirling's equation gave an expression for the entropy that is crucial for explaining the radiation of black bodies.

Problem j Estimate $n!$ when n is equal to Avogrado's number. Hint: take the logarithm of (24.30).

24.3 Method of steepest descent

In this section we treat the method of steepest descent. This method for the asymptotic evaluation of integrals is based on the idea that many functions in the complex plane have a stationary point. There is one direction at that stationary point in which the function decreases rapidly, and there is an orthogonal direction in which the function increases rapidly. By deforming the integration path so that it goes through

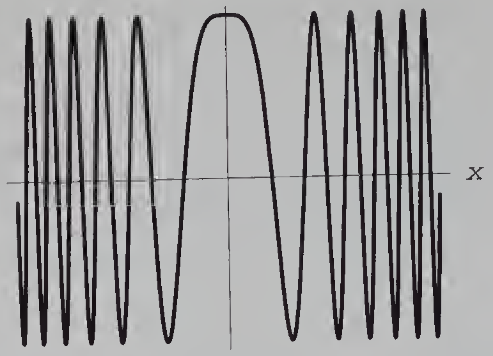

Fig. 24.3 The function $\Re e\left(\exp\left(ix^2\right)\right)$ for $-6 < x < 6$. The vertical axis runs from -1 to 1.

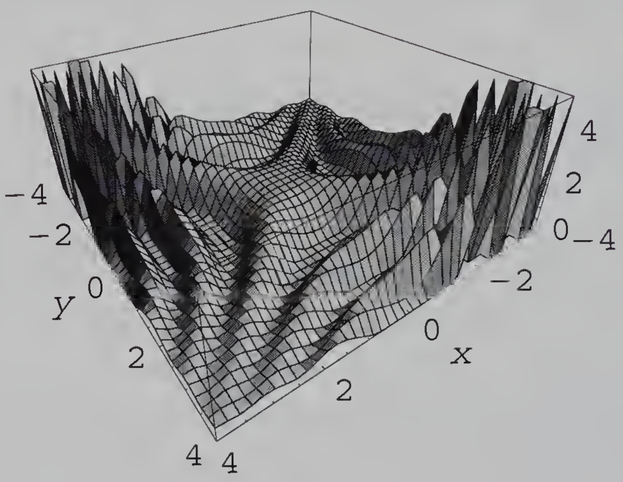

Fig. 24.4 The function $\Re e\left(\exp\left(iz^2\right)\right)$ in the complex plane. For clarity the exponential growth has been modified by plotting the function $e^{-0.1\times 2xy}\cos(x^2 - y^2)$.

the stationary point in the direction in which the function decreases, one can evaluate the integral asymptotically. As an example we apply this idea to the following integral

$$I = \int_{-\infty}^{\infty} e^{iax^2}\,dx, \qquad \text{with } a > 0. \tag{24.32}$$

Let us first investigate the behavior of the integrand $\exp\left(iz^2\right)$ in the complex plane with $z = x + iy$ for the special case $a = 1$. The real part of the integrand along the real axis is shown in Figure 24.3. Along this axis, the integrand is a rapidly oscillating function, except near the point $x = 0$ where the function is stationary. The real part of $\exp\left(iz^2\right)$ in the complex plane is shown in Figure 24.4. It can be seen from this figure that along the real axis and along the imaginary axis, the function is oscillatory. Along the line $y = x$ the function has a *maximum* at the

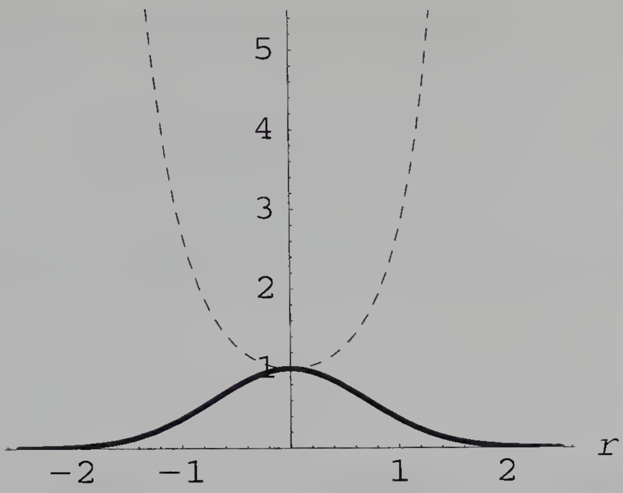

Fig. 24.5 The function $\Re e\left(\exp\left(iz^2\right)\right)$ in the complex plane along the lines $x = y$ (solid line) and $x = -y$ (dashed line).

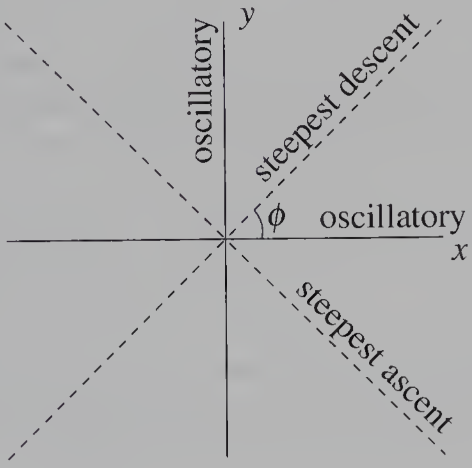

Fig. 24.6 The behaviour of the function e^{iz^2} in the complex plane.

origin, while along the line $y = -x$ the function has a *minimum* at the origin. The behavior of $\Re e\left(\exp\left(iz^2\right)\right)$ along these lines is shown in Figure 24.5.

Problem a Compute $\Re e\left(\exp\left(iz^2\right)\right)$ along the x-axis, the y-axis, and the lines $y = \pm x$ and show that this function has a behavior as shown in Figure 24.6. A direction of steepest descent means that in that direction the function *decreases* from its stationary point, while a direction of steepest ascent means that in that direction a function *increases* from its stationary point.

There is a good reason why there are steepest descent and steepest ascent directions that are perpendicular. As shown in Section 16.1, the real and imaginary parts of an analytic function satisfy Laplace's equation: $\nabla^2 f = 0$. Let us momentarily

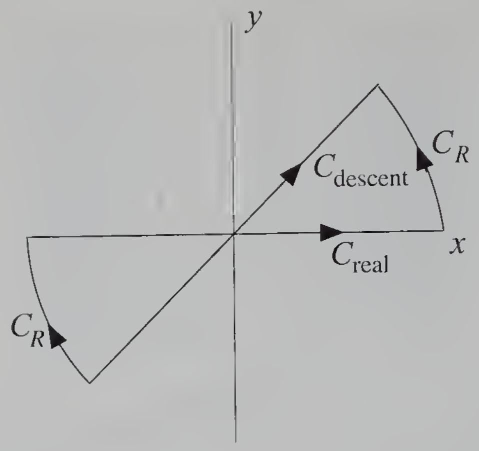

Fig. 24.7 Definition of the integrations paths for the steepest descent integration.

assume that the x-axis at the stationary point is aligned with the steepest descent direction. This means that the stationary point is a maximum, and according to Section 10.1 this means that $\partial^2 f/\partial x^2 < 0$. Since f satisfies Laplace's equation $\partial^2 f/\partial x^2 + \partial^2 f/\partial y^2 = 0$, this means that at the stationary point $\partial^2 f/\partial y^2 > 0$, hence the stationary point is a minimum in the y-direction. The stationary point is a *saddle point*; the function decreases in one direction while it increases in another direction. This can clearly be seen in Figure 24.4. This saddle point behavior forms the basis of Earnshaw's theorem that was treated in Section 10.5.

We now use the steepest descent path $y = x$ to evaluate the integral (24.32). In order to do this, we deform the integration path C_{real} into the steepest descent path $C_{descent}$ as shown in Figure 24.7. In doing so we need to close the contour with two circle segments C_R with a radius R that goes to infinity.

Problem b Use the fact that e^{iz^2} is analytic within the closed contour of Figure 24.7 to show that

$$\int_{-\infty}^{\infty} e^{iax^2} dx + \int_{C_R} e^{iaz^2} dz - \int_{C_{descent}} e^{iaz^2} dz = 0. \tag{24.33}$$

Problem c At the contour C_R, $z = R\, e^{i\varphi}$ with $0 < \varphi < \pi/4$ or $\pi < \varphi < 5\pi/4$. Show that along the contour C_R:

$$\left| e^{iaz^2} \right| = e^{-aR^2 \sin 2\varphi}, \tag{24.34}$$

and that for the employed values of φ this function decays exponentially as $R \to \infty$.

This result implies that the contribution of the contours C_R vanishes as $R \to \infty$, so that according to (24.33):

$$\int_{-\infty}^{\infty} e^{iax^2} dx = \int_{C_{descent}} e^{iaz^2} dz. \tag{24.35}$$

Problem d Evaluate the integral on the right-hand side by substituting $z = ue^{i\pi/4}$ and use (24.28) to show that

$$\int_{-\infty}^{\infty} e^{iax^2} dx = \sqrt{\frac{\pi}{a}}\, e^{i\pi/4} \qquad (a > 0). \tag{24.36}$$

Problem e Show that this result can also be obtained by simply substituting $b = -ia$ into the integral (24.28). This substitution does not address any convergence issues of the resulting integral. However, it is a simple trick to memorize (24.36).

Problem f Repeat the analysis that leads to problem d for $a < 0$. Show that in this case the steepest descent direction is along the line $y = -x$ and that

$$\int_{-\infty}^{\infty} e^{iax^2} dx = \sqrt{\frac{\pi}{|a|}}\, e^{-i\pi/4} \qquad (a < 0). \tag{24.37}$$

Note that the expressions (24.36) and (24.37) can be summarized by the single expression

$$\int_{-\infty}^{\infty} e^{iax^2} dx = \sqrt{\frac{\pi}{|a|}} e^{i\pi/4\, sgn(a)}, \tag{24.38}$$

with

$$sgn(a) \equiv \begin{cases} +1 \text{ when } a > 0 \\ -1 \text{ when } a < 0 \end{cases}. \tag{24.39}$$

The integration in this section was over a single variable. Bleistein and Handelsman [18] show how this analysis can be extended to multidimensional integrals:

$$\int \exp(i\mathbf{x} \cdot \mathbf{A} \cdot \mathbf{x}) d^N\mathbf{x} = \sqrt{\frac{\pi^N}{|\det \mathbf{A}|}}\, e^{i\pi/4\, sgn(\mathbf{A})}, \tag{24.40}$$

where $sgn(\mathbf{A})$ = *number of positive eigenvalues of* **A** *– number of negative eigenvalues of* **A**.

The advantage of using expression (24.38) is that we do not need to know the steepest descent path once we know this integral. The integral (24.38) forms the basis of the method of stationary phase that will be treated in the next section. Since the method of stationary phase uses the original integration path along the real axis, one can say that *the method of stationary phase is the steepest descent method for those that are too lazy to deform contours in the complex plane.*

24.4 Group velocity and the method of stationary phase

Before treating the method of stationary phase, let us briefly examine a common explanation of the concept of group velocity. Suppose a wave consists of two propagating waves of equal amplitudes with frequencies ω_1 and ω_2 that are close:

$$p(x,t) = \cos(k(\omega_1)x - \omega_1 t) + \cos(k(\omega_2)x - \omega_2 t) \tag{24.41}$$

Note that the wave number is a function of frequency: $k = k(\omega)$. In the following we use

$$\omega_1 = \omega_0 - \Delta \qquad \text{and} \qquad \omega_2 = \omega_0 + \Delta, \tag{24.42}$$

so that ω_0 is the center frequency and 2Δ the frequency separation.

Problem a Use a first-order Taylor expansion of $k(\omega)$ around the point ω_0 to show that

$$\begin{aligned}\cos(k(\omega_1)x - \omega_1 t) = {} & \cos(k_0 x - \omega_0 t)\cos\left(\Delta\left(\frac{\partial k}{\partial \omega}x - t\right)\right)\\ & + \sin(k_0 x - \omega_0 t)\sin\left(\Delta\left(\frac{\partial k}{\partial \omega}x - t\right)\right),\end{aligned} \tag{24.43}$$

where $k_0 = k(\omega_0)$ and $\partial k/\partial \omega$ is evaluated at ω_0.

Problem b Apply the same analysis to the last term of (24.41) and derive that

$$p(x,t) = 2\cos\left(\omega_0(t - \frac{x}{c})\right)\cos\left(\Delta\left(t - \frac{x}{U}\right)\right), \tag{24.44}$$

with

$$c = \frac{\omega}{k} \qquad \text{and} \qquad U = \frac{\partial \omega}{\partial k}. \tag{24.45}$$

Since $\Delta \ll \omega_0$, the total wave field (24.44) consists of a wave $\cos(\omega_0(t - (x/c)))$ that propagates with a velocity c that is modulated by a slowly varying amplitude variation $\cos(\Delta(t - x/U))$ that propagates with a velocity U. The velocity U of the amplitude variation is called the *group velocity*, while the velocity c with which the carrier wave of frequency ω_0 propagates is called the *phase velocity*.

Even though this analysis is relatively simple, it is not very realistic. A propagating wave rarely consists of two frequency components that have equal amplitudes. In general, a wave consists of the superposition of all frequency components within a certain frequency band, and the amplitude spectrum $A(\omega)$ in general is not a constant. In the frequency domain such a wave is given by $A(\omega)\exp(ik(\omega)x)$, and after a Fourier transform the wave is in the time domain given by

$$p(x,t) = \int_{-\infty}^{\infty} A(\omega) e^{i(k(\omega)x-\omega t)} d\omega. \tag{24.46}$$

In the following we assume that the amplitude $A(\omega)$ varies slowly with frequency compared to the phase

$$\psi(\omega) = k(\omega)x - \omega t. \tag{24.47}$$

In general, the phase is a rapidly oscillating function of frequency. We argued in the previous section that the dominant contribution to the integral $\int_{-\infty}^{\infty} e^{iax^2} dx$ came from the saddle point $x = 0$. In Figure 24.3 the saddle point $x = 0$ corresponds to the point where the function ceases to oscillate. This corresponds to the point where the phase in the exponent is stationary, that is this is the point where the phase is either a maximum or a minimum. For the function e^{iax^2} the phase is given by ax^2, the phase is stationary when $d\left(ax^2\right)/dx = 0$; this implies that $x = 0$, which is indeed the point of stationary phase in Figure 24.3.

Problem c Show that the phase (24.47) is stationary when

$$\frac{\partial k}{\partial \omega} x - t = 0. \tag{24.48}$$

With the definition (24.45) for the group velocity this result can also be written as

$$U(\omega_0) = \frac{x}{t}. \tag{24.49}$$

Before we proceed, let us briefly reflect on this result. The Fourier integral (24.46) states that all frequencies contribute to the wave field. However, for given values of x and t the dominant contribution to this frequency integral comes from the frequency ω_0 for which the phase of the integrand is stationary. This frequency is defined by expression (24.49). When the group velocity is known as a function of frequency, this expression implicitly defines ω_0. Note that this frequency depends on the location x and the time t for which we want to compute the wave field. The right-hand side x/t is the distance x covered in a time t, the principle of stationary phase simply states that this distance is covered with a velocity given by the group velocity.

We approximate the integral (24.46) here with the method of stationary phase. In this approximation we assume that:

- The amplitude $A(\omega)$ varies slowly with ω, so that we can replace the amplitude by its value at the stationary point:
$$A(\omega) \approx A(\omega_0). \tag{24.50}$$
- For the phase we use a second order Taylor expansion around the stationary point:
$$\psi(\omega) \approx \psi(\omega_0) + \frac{1}{2}\frac{\partial^2 \psi}{\partial \omega^2}(\omega - \omega_0)^2, \tag{24.51}$$
where the second derivative is evaluated at the stationary point ω_0. Note that the first derivative does not appear in this Taylor expansion because at the stationary point we have by definition that $\partial\psi/\partial\omega = 0$.
- Insert these approximations into (24.46) and evaluate the remaining integral with expression (24.38).

Problem d The first step in this approach is trivial. In the second step we need $\partial^2\psi/\partial\omega^2$ at the stationary point. Compute this derivative and use (24.45) to show that

$$\frac{\partial^2 \psi}{\partial \omega^2} = \frac{\partial^2 k}{\partial \omega^2}x = -\frac{1}{U^2}\frac{\partial U}{\partial \omega}x, \tag{24.52}$$

where all quantities are evaluated at frequency ω_0.

Problem e Insert (24.50)–(24.52) into the integral (24.46) and solve the remaining integral with (24.38) to show that

$$p(x,t) \approx A(\omega_0)e^{i(k_0 x - \omega_0 t)}\sqrt{\frac{2\pi}{|\partial U/\partial \omega|\, x}}\; U(\omega_0)e^{-i\pi/4\, sgn(\partial U/\partial\omega)}. \tag{24.53}$$

This expression generalizes (24.44) for a wave with an arbitrary amplitude spectrum $A(\omega)$. The wave (24.53) consists of a carrier wave $\exp(i(k_0x - \omega_0 t))$ with amplitude $A(\omega_0)$. The frequency ω_0 is determined by the condition that the group velocity $U(\omega_0)$ is the velocity needed to cover the distance x in time t. The amplitude of the wave is also determined by the term $1/\sqrt{|\partial U/\partial \omega|\, x}$. Let us first consider the x-dependence. The wave we consider here has a group velocity that depends on frequency. This means that the different frequency components travel with a different velocity. Because of this, the amplitude of the wave is reduced as the wave propagates, because the different frequency components are spread-out over space during the propagation. This phenomenon is called *dispersion.* The decay of the amplitude with propagation distance is given by the term $1/\sqrt{x}$. Suppose that the group velocity depends weakly on frequency. In that case $\partial U/\partial\omega$ is small. In that situation, many different frequency components interfere constructively in the Fourier integral. This leads to a large amplitude that is described by the term $1/\sqrt{|\partial U/\partial\omega|}$. An extensive treatment of the concept of group velocity is given by Brillouin [21].

An example of wave dispersion can be seen in Figure 20.11 which shows the vertical component of the ground motion recorded in Belarus after an earthquake at Jan-Mayen Island. Between $t = 650$ s and $t = 900$ s a dispersive wave arrives. This is the Rayleigh wave, an elastic wave that is guided along the Earth's surface. The frequency content at early times ($t = 700$ s) is lower than at later times (say $t = 800$ s). This dispersion is caused by the fact that the group velocity of the Rayleigh wave decreases with increasing frequency for the frequencies that contribute to the seismogram in Figure 20.11. This causes the low-frequency components to arrive earlier than the high-frequency components.

It may occur that the group velocity does not depend on frequency. According to (24.52) this means that $\partial^2\psi/\partial\omega^2 = 0$ and the stationary phase integral (24.53) is infinite. In that case the Taylor expansion (24.51) must be replaced by the third order Taylor expansion

$$\psi(\omega) \approx \psi(\omega_0) + \frac{1}{6}\frac{\partial^3\psi}{\partial\omega^3}(\omega - \omega_0)^3. \tag{24.54}$$

The resulting integral can be integrated to give an Airy function [14]. The corresponding wave arrivals are called *Airy phases* [114]; these waves have, in general, a strong amplitude because many different frequency components interfere constructively.

24.5 Asymptotic behavior of the Bessel function $J_0(x)$

In Section 20.5 we showed that for $x \gg n$ the Bessel function $J_n(x)$ behaves as a decaying cosine:

$$J_n(x) \approx \frac{A}{\sqrt{x}}\cos(x + \varphi). \tag{24.55}$$

This relation followed from an analysis of the differential equation of the Bessel function for large values of x. The analysis of Section 20.5 did not tell us what the constants A and φ are. Here we find these constants for $J_0(x)$ by applying a stationary phase analysis to the following integral representation of the Bessel function:

$$J_0(x) = \frac{1}{\pi} \int_0^\pi \cos(x \sin\varphi)\, d\varphi. \tag{24.56}$$

Before we solve this integral in the stationary phase approximation, let us verify that this is a valid representation of the Bessel function. In order to do this, we need to check if the representation (24.56) satisfies the Bessel equation (20.16), with the following initial conditions:

$$J_0(x=0) = 1, \qquad \frac{dJ_0}{dx}(x=0) = 0. \tag{24.57}$$

Problem a Show that

$$\frac{dJ_0}{dx} = \frac{-1}{\pi} \int_0^\pi \sin(x \sin\varphi) \sin\varphi\, d\varphi, \tag{24.58}$$

and

$$\frac{d^2 J_0}{dx^2} = \frac{-1}{\pi} \int_0^\pi \cos(x \sin\varphi) \sin^2\varphi\, d\varphi. \tag{24.59}$$

Using the identity $\sin^2\varphi = 1 - \cos^2\varphi$, the last integral can be written as $-\pi^{-1} \int_0^\pi \cos(x \sin\varphi)\, d\varphi + \pi^{-1} \int_0^\pi \cos(x \sin\varphi) \cos^2\varphi\, d\varphi$. In the last term we can use that

$$\cos(x \sin\varphi) = \frac{1}{x\cos\varphi} \frac{d \sin(x \sin\varphi)}{d\varphi}. \tag{24.60}$$

Problem b Insert these results into (24.59), and use an integration by parts of the last term to derive that

$$\frac{d^2 J_0}{dx^2} = \frac{-1}{\pi} \int_0^\pi \cos(x \sin\varphi)\, d\varphi + \frac{1}{\pi x} \int_0^\pi \sin(x \sin\varphi) \sin\varphi\, d\varphi. \tag{24.61}$$

Problem c Use this result with (24.56) and (24.58) to show that the representation (24.56) satisfies the Bessel equation (20.16).

Problem d We next check if $J_0(x)$ indeed satisfies the initial conditions (24.57). Apply the Taylor approximation technique of Section 24.1 to the integral (24.56) and show that for small values of x:

$$J_0(x) = 1 + 0 \cdot x + O(x^2). \tag{24.62}$$

This result implies that the initial conditions (24.57) are satisfied, so that (24.56) indeed is a valid representation of the Bessel function $J_0(x)$.

In order to apply a stationary phase analysis for $x \gg 1$ to (24.56) we first rewrite it in the following form:

$$J_0(x) = \frac{1}{\pi} \Re e \int_0^{\pi} e^{ix\psi(\varphi)} d\varphi, \tag{24.63}$$

with

$$\psi(\varphi) = \sin \varphi. \tag{24.64}$$

Problem e Show that the phase ψ is stationary for $\varphi = \pi/2$ and that at that point $d^2\psi/d\varphi^2 = -1$.

Problem f Use a second order Taylor expansion around the stationary point to rewrite (24.63) as

$$J_0(x) \approx \frac{1}{\pi} \Re e \, e^{ix} \int_0^{\pi} \exp\left(-\frac{ix}{2} (\varphi - \frac{\pi}{2})^2 \right) d\varphi. \tag{24.65}$$

The integrand has the behavior shown in Figure 24.3; away from the stationary point the integrand fluctuates rapidly and the positive contributions are canceled by the negative contributions so that the dominant contribution to the integral comes from a region near the stationary point.

Problem g We can estimate the region of stationary phase by finding the values of φ for which the real part of the integrand of (24.65) vanishes. These points correspond to the zero crossing in Figure 24.3 closest to the stationary point. Show that the corresponding values of φ are given by

$$\varphi = \frac{\pi}{2} \pm \sqrt{\frac{\pi}{x}}. \tag{24.66}$$

For $x \gg 1$ these values of φ are close to the stationary point. Since the dominant contribution to the stationary phase integral (24.65) comes from the region close to the stationary point, we can extend the integration interval to infinity:

$$J_0(x) \approx \frac{1}{\pi} \Re e \, e^{ix} \int_{-\infty}^{\infty} \exp\left(-\frac{ix}{2} (\varphi - \frac{\pi}{2})^2 \right) d\varphi. \tag{24.67}$$

Problem h Use (24.38) to show that

$$J_0(x) \approx \frac{1}{\pi}\Re e \sqrt{\frac{2\pi}{x}} e^{i(x-\pi/4)}, \tag{24.68}$$

and that

$$J_0(x) \approx \sqrt{\frac{2}{\pi x}} \cos\left(x - \frac{\pi}{4}\right). \tag{24.69}$$

Bender and Orszag [14] derive the same result with a steepest descent analysis of the integral (24.56). Their analysis is much more complicated than the stationary phase analysis of this section.

24.6 Image source

In this section we consider waves in three dimensions that are reflected from a plane $z = 0$. The geometry of this problem is shown in Figure 24.8. A source at $\mathbf{r}_s$ emits waves that propagate to a reflection point $\mathbf{r}$ in the reflection plane. The waves are reflected at that point with reflection coefficient $R(x, y)$, and then propagate to the receiver $\mathbf{r}_r$. We assume here that the reflection coefficient $R(x, y)$ varies slowly with the position on the reflecting plane. The total reflected wave follows by an integration over the reflection surface. In this section we use a coordinate system as shown in Figure 24.8. The reflection plane is defined by $z = 0$, and the y-axis of the coordinate system is aligned with the source–receiver positions in such a way

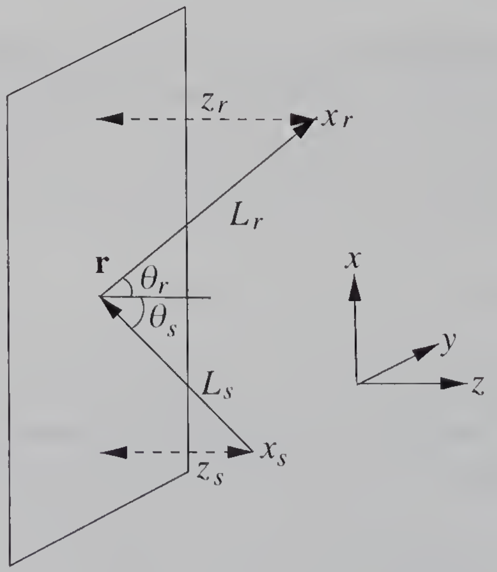

Fig. 24.8 Definition of the geometric variables for the waves that reflect in a plane.

that:

$$\mathbf{r}_s = \begin{pmatrix} x_s \\ 0 \\ z_s \end{pmatrix}, \quad \mathbf{r} = \begin{pmatrix} x \\ y \\ 0 \end{pmatrix}, \quad \mathbf{r}_r = \begin{pmatrix} x_r \\ 0 \\ z_r \end{pmatrix}. \tag{24.70}$$

The propagation of the waves from the source to the reflection point, and from the reflection point to the receiver is given by the Green's function (19.37) for scalar waves in a 3D homogeneous medium. The total wave field is given by the integral over the reflecting plane, so that

$$p(\mathbf{r}_r) = \iint\limits_{-\infty}^{\infty} G(|\mathbf{r}_r - \mathbf{r}|)R(x, y)G(|\mathbf{r} - \mathbf{r}_s|)\, dxdy, \tag{24.71}$$

with G given by expression (19.37).

Problem a Show that this integral can be written as

$$p(\mathbf{r}_r) = \iint\limits_{-\infty}^{\infty} A(x, y)R(x, y)e^{ik\psi(x,y)}dxdy, \tag{24.72}$$

with

$$A(x, y) = \frac{1}{(4\pi)^2}\left((x_r - x)^2 + y^2 + z_r^2\right)^{-1/2}\left((x_s - x)^2 + y^2 + z_s^2\right)^{-1/2}, \tag{24.73}$$

and

$$\psi(x, y) = \sqrt{(x_r - x)^2 + y^2 + z_r^2} + \sqrt{(x_s - x)^2 + y^2 + z_s^2} \tag{24.74}$$

Problem b We will solve the integral (24.72) in the stationary phase approximation. In order to do this we evaluate both the x- and the y-integral with a stationary phase approximation. Before we can do this, we first need to find the stationary point in the x, y-plane. In order to find this point, show first that

$$\frac{\partial \psi}{\partial x} = \frac{x - x_r}{\sqrt{(x_r - x)^2 + y^2 + z_r^2}} + \frac{x - x_s}{\sqrt{(x_s - x)^2 + y^2 + z_s^2}}, \tag{24.75}$$

and

$$\frac{\partial \psi}{\partial y} = \frac{y}{\sqrt{(x_r - x)^2 + y^2 + z_r^2}} + \frac{y}{\sqrt{(x_s - x)^2 + y^2 + z_s^2}}. \tag{24.76}$$

Problem c At the stationary point, both derivatives vanish; show that the stationary point is given by

$$\frac{x - x_r}{L_r} + \frac{x - x_s}{L_s} = 0 \qquad \text{and} \qquad y = 0, \tag{24.77}$$

with L_r and L_s defined in Figure 24.8.

Let us first analyze this stationary phase condition. The condition $y = 0$ states that the line from the source to the reflection point and the line from the reflection point to the receiver lie in the same plane. The phase is thus stationary for *in-plane* reflection. The first equality in (24.77) states for which point in that plane the phase is stationary.

Problem d Use the angles θ_s and θ_r defined in Figure 24.8 to show that the point of stationary phase as defined by the first equality in (24.77) satisfies

$$\theta_s = \theta_r . \tag{24.78}$$

This identity implies that at the point of stationary phase the angles of the incoming and the outgoing waves are equal. If the waves could be described by rays, the reflection point would be defined by (24.78) and also the condition $y = 0$. The condition of stationary phase thus gives the same reflection point as would be given by ray theory. A similar condition holds when the incoming and outgoing waves travel at a different velocity. Snieder [97] shows that for the reflection of surface waves where the incoming surface wave mode travels with velocity c_{in} and the outgoing surface wave mode with velocity c_{out} the point of stationary phase satisfies

$$\frac{\sin\theta_{in}}{c_{in}} = \frac{\sin\theta_{out}}{c_{out}} . \tag{24.79}$$

This is nothing but Snell's law [3, 53] for mode-converted waves.

The condition of stationarity thus gives us the location of the ray-geometric reflection point. Let is now compute the reflected wavefield. In order to evaluate (24.72) in the stationary phase approximation we need the second derivatives of the phase.

Problem e Show for an arbitrary point (x, y) that

$$\frac{\partial^2 \psi}{\partial x^2} = \frac{y^2 + z_r^2}{\left((x_r - x)^2 + y^2 + z_r^2\right)^{3/2}} + \frac{y^2 + z_s^2}{\left((x_s - x)^2 + y^2 + z_s^2\right)^{3/2}}, \tag{24.80}$$

and

$$\frac{\partial^2 \psi}{\partial y^2} = \frac{(x_r - x)^2 + z_r^2}{\left((x_r - x)^2 + y^2 + z_r^2\right)^{3/2}} + \frac{(x_s - x)^2 + z_s^2}{\left((x_s - x)^2 + y^2 + z_s^2\right)^{3/2}}. \tag{24.81}$$

Problem f Show that at the stationary point:

$$\frac{\partial^2 \psi}{\partial x^2} = \cos^2\theta \left(\frac{1}{L_r} + \frac{1}{L_s}\right), \tag{24.82}$$

and

$$\frac{\partial^2 \psi}{\partial y^2} = \left(\frac{1}{L_r} + \frac{1}{L_s}\right), \tag{24.83}$$

where $\theta = \theta_r = \theta_s$ at the reflection point, and L_s and L_r are the distance from the reflection point to the source and receiver, respectively.

Problem g Show that near the reflection point the phase is approximately given by

$$\psi(x, y) \approx L_r + L_s + \frac{1}{2}\cos^2\theta \left(\frac{1}{L_r} + \frac{1}{L_s}\right)(x - x_{refl})^2 + \frac{1}{2}\left(\frac{1}{L_r} + \frac{1}{L_s}\right)(y - y_{refl})^2, \tag{24.84}$$

with (x_{refl}, y_{refl}) the coordinates of the reflection point.

Problem h Show that the amplitude A near the reflection point is given by

$$A(x, y) = \frac{1}{(4\pi)^2 L_r L_s}. \tag{24.85}$$

Problem g Evaluate the x and y integral in (24.72) each with the stationary phase approximation, and show that in this approximation

$$p(\mathbf{r}_r) = r\frac{\exp(ik(L_r + L_s))}{4\pi(L_r + L_s)}, \tag{24.86}$$

with

$$r = \frac{iR}{2k\cos\theta}. \tag{24.87}$$

In this expression R is the reflection coefficient at the reflection point at an angle of incidence θ.

Apart from a minus sign, the term $\exp(ik(L_r + L_s))/4\pi(L_r + L_s) = -G(L_r + L_s)$ is the Green's function for a distance $L_r + L_s$. The wavefield (24.86)

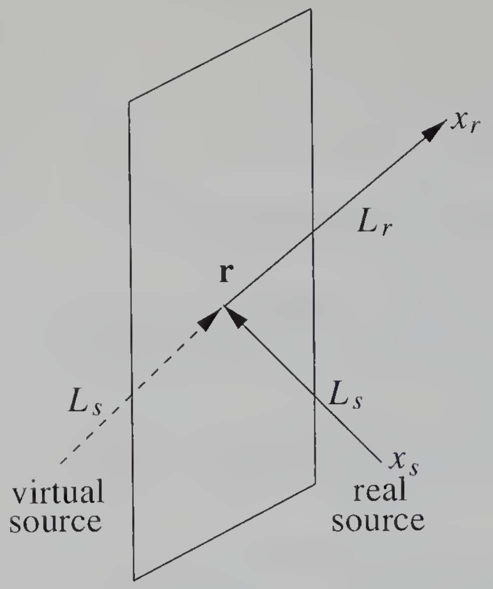

Fig. 24.9 The position of a virtual source that is the mirror image of the real source.

can therefore be interpreted as the wavefield generated by a virtual source that is the mirror image of the original source in the reflector plane, see Figure 24.9. This corresponding Green's function $G(L_r + L_s)$ is multiplied in (24.86) with an effective reflection coefficient r that accounts for the net reflectivity of the reflection plane. The stationary phase approximation thus not only accounts for the ray-geometric reflection point, it also describes how the original source can be replaced by an image source at the other side of the reflection point.

25

Variational calculus

25.1 Designing a can

As an introduction to variational calculus we consider in this section the problem of designing a can. Suppose you want to make the can with a given amount of metal that has the largest content. A squat can with a small height and a large radius has a small volume, but a tall can with a large height but a small radius cannot contain much either. Clearly there is a specific shape of the can that maximizes the volume for a given surface area. Consider the can shown in Figure 25.1, the top and bottom each having a surface area S, the height denoted by h, and the circumference of the can denoted by C. The total surface area A is twice the surface area of the bottom, plus the surface area hC of the side, so that

$$A = 2S + hC. \tag{25.1}$$

The volume V of the can is the product of the surface area of the bottom times the height:

$$V = hS. \tag{25.2}$$

Problem a We want to eliminate the height h from the problem. Use expression (25.1) to show that

$$h = \frac{(A - 2S)}{C}, \tag{25.3}$$

and use this result to show that

$$V = \frac{(A - 2S)\,S}{C}. \tag{25.4}$$

Problem b For the moment we assume that the can has the shape of cylinder with radius r. Express the surface area S and the circumference C in this radius,

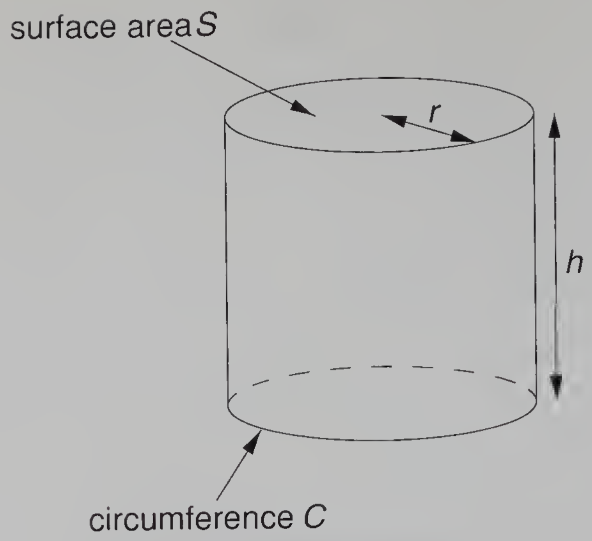

Fig. 25.1 Definition of the geometric variables of a can.

and show that the volume is given by

$$V(r) = \frac{r}{2}(A - 2\pi r^2). \tag{25.5}$$

The volume of the can is now given as a function of r. The value of the radius that maximizes the volume can be found by requiring that the derivative of the volume with respect to the radius vanishes:

$$\frac{dV}{dr} = 0. \tag{25.6}$$

Problem c Differentiate expression (25.5) to show that the radius that gives the largest volume is given by

$$r = \sqrt{\frac{A}{6\pi}}. \tag{25.7}$$

Problem d Insert this value into expression (25.3) while using the appropriate expressions for the surface and the circumference of a circle to show that

$$h = \sqrt{\frac{2A}{3\pi}}, \tag{25.8}$$

and that the optimal ratio of the height and the radius is given by

$$\frac{h}{r} = 2. \tag{25.9}$$

This expression defines the shape of the can. A can whose height is equal to its diameter $2r$ holds the largest volume for a given surface area. The treatment of this section is called an *optimization problem*. In such a problem some quantity is being optimized. Often a straightforward differentiation such as in expression (25.6) gives the optimum value of a parameter. Optimization problems arise naturally in many design problems, but also in economic problems and other situations where a strategy needs to be formulated.

25.2 Why are cans round?

When my daughter noticed I entered this title, she thought this was an utterly stupid question. However, is it obvious that a can with a cylindrical shape contains the largest volume? Why should not the cross section of a can be ellipsoidal or square? In the previous section we optimized the shape of a cylindrical can by the simple differentiation (25.6) with respect to the single parameter r. However, for an arbitrary shape of the can, the radius r is a function of the angle φ as shown in Figure 25.2. In that situation the radius is not a single parameter, and the problem is to find the *function* $r(\varphi)$ that maximizes the volume of the can. This is a different problem that cannot be solved by simple differentiation, because the radius $r(\varphi)$ is a function rather than a single parameter.

It follows from expression (25.4) that for a given value of the surface S, the volume is largest when the circumference C is smallest. The problem of finding the function $r(\varphi)$ that maximizes the volume thus is equivalent to finding the function $r(\varphi)$ that minimizes the circumference C for a given surface area S. Let us first express the circumference C in the function $r(\varphi)$.

Problem a The radius vector **r** is given by

$$\mathbf{r} = \begin{pmatrix} r(\varphi)\cos\varphi \\ r(\varphi)\cos\varphi \end{pmatrix}. \tag{25.10}$$

Show that

$$\frac{d\mathbf{r}}{d\varphi} = \begin{pmatrix} \dot{r}(\varphi)\cos\varphi - r(\varphi)\sin\varphi \\ \dot{r}(\varphi)\sin\varphi + r(\varphi)\cos\varphi \end{pmatrix}, \tag{25.11}$$

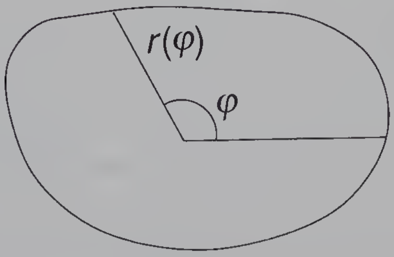

Fig. 25.2 The radius of a can with an arbitrary shape as a function of the angle φ.

with

$$\dot{r} \equiv \frac{dr}{d\varphi}. \tag{25.12}$$

Problem b An increment ds along the circumference that corresponds with an increment $d\varphi$ satisfies $(ds/d\varphi)^2 = \left(\frac{d\mathbf{r}}{d\varphi} \cdot \frac{d\mathbf{r}}{d\varphi} \right)$. Use this relation and expression (25.11) to derive that

$$ds = \sqrt{\dot{r}^2 + r^2}\, d\varphi. \tag{25.13}$$

Integrating this expression over a full circle gives the circumference:

$$C = \int_0^{2\pi} \sqrt{\dot{r}^2 + r^2}\, d\varphi. \tag{25.14}$$

A function constitutes a recipe that maps one variable onto another variable. An example is the function $V(r)$ in expression (25.5) that gives the volume of a cylindrical can for a given value of the radius. The integral (25.14) is different. The quantity r in the right-hand side is not a single variable, since r now is a function itself of the angle φ. In fact, expression (25.14) is a relation that maps a function $r(\varphi)$ onto a single variable; in this case the circumference C. In order to contrast this from the normal behavior of a function, the circumference C is called a *functional* of the function $r(\varphi)$. Usually this dependence in indicated with the notation $C[r]$. In general, a functional is a number that is attached to a function. For example, the mean $\overline{f}$ of a function $f(x)$ is a functional of $f(x)$. In general, many different functionals can be attached to a function. In our problem the functional of interest is the circumference given in expression (25.14).

Here we want to find the function $r(\varphi)$ that minimizes the circumference for a given surface area. Let us find the solution by examining expression (25.14). The quantity $\dot{r}^2$ is always positive. This means that the circumference C is smallest when $\dot{r} = 0$, because this is the value that minimizes $\dot{r}^2$. The condition $\dot{r} = 0$ implies that $r = constant$, which in turn states that the base of the can is a circle, hence the can that holds the largest volume for a given surface area is indeed cylindrical.

In this particular example we could see the solution rather easily. However, the reasoning used is not very rigorous, and in more complex problems the solution cannot be seen in a simple way. In this chapter we develop systematic tools to find the maximum or minium of a functional. The mathematics that is involved is called *variational calculus*. We return to the problem of minimizing the circumference in Section 25.9. In this chapter we cover the basic principles of variational calculus. Smith [95] gives a comprehensive and clear overview of this topic.

By the way, why do not cans have a hexagonal shape? In this way they can be stacked without having voids between the cans, and they are probably stronger. This shows that in design there often are different design criteria that must be balanced.

25.3 Shortest distance between two points

It seems obvious that the shortest distance between two points is a straight line, but a proof of this property is not trivial. In this section we give such a proof because it offers a prototype of the approach taken in variational calculus. Consider the situation shown in Figure 25.3 where we seek the shortest line that joins the points (x_1, y_1) and (x_2, y_2). In the following we take x as the independent parameter and seek the function $y(x)$ that goes through the two endpoints and gives the curve with the smallest length.

Problem a In order to solve this problem we first need to compute the length of the curve for a given function $y(x)$. A point on the curve has coordinates

$$\mathbf{r} = \begin{pmatrix} x \\ y(x) \end{pmatrix} . \tag{25.15}$$

Compute the increment $d\mathbf{r}$ due to an increment dx, and use the relation $ds^2 = (d\mathbf{r} \cdot d\mathbf{r})$ to show that

$$ds = \sqrt{1 + \dot{y}^2}\, dx , \tag{25.16}$$

where

$$\dot{y} \equiv \frac{dy}{dx} . \tag{25.17}$$

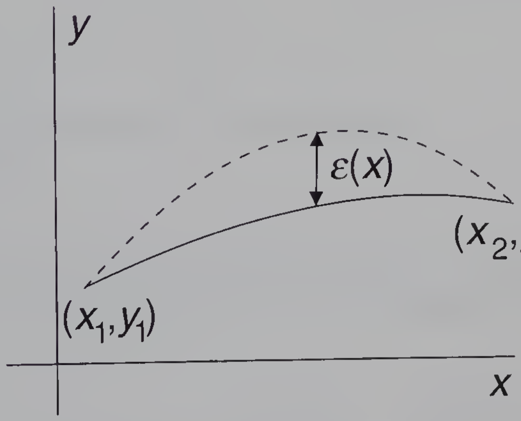

Fig. 25.3 A curve $y(x)$ that joins two points (solid line), the perturbation $\varepsilon(x)$, and the perturbed curve (dashed line).

The total length L of the curve follows by integrating expression (25.16) from x_1 to x_2:

$$L[y] = \int_{x_1}^{x_2} \sqrt{1 + \dot{y}^2}\, dx. \tag{25.18}$$

The length $L[y]$ is a functional of the function $y(x)$.

How do we find the function $y(x)$ so that $L[y]$ is minimized? Consider first the simpler problem that we want to minimize a function $f(x)$ as a function of the variable x. The function has an extremum at x_{extr} when the first derivative vanishes: $df/dx = 0$ at x_{extr}. Let us perturb the value of x around the extremum with a value ε. A first-order Taylor expansion then gives

$$f(x_{extr} + \varepsilon) = f(x_{extr}) + \frac{df}{dx}(x_{extr})\varepsilon + O(\varepsilon^2) = f(x_{extr}) + O(\varepsilon^2). \tag{25.19}$$

The last identity follows from the fact that the function f has an extremum at x_{extr} and its first derivative therefore vanishes at that point. In words, expression (25.19) can be stated as follows: the point x_{extr} is an extremum of the function f when a perturbation of x around the extremal point does not lead to first-order to a change in the function.

We can use a similar reasoning for the minimization of the length L in expression (25.18). The only difference is that $L[y]$ is a functional of the function y, whereas $f(x)$ is a function of the variable x. Suppose we have a given curve $y(x)$, as shown by the solid line in Figure 25.3. Let us perturb that curve with a perturbation $\varepsilon(x)$ so that we get a new curve. When the curve $y(x)$ minimizes the length L, then the perturbation $\varepsilon(x)$ should to first order not change that length. This requirement for the minimization of L is analogous to the last identity in expression (25.19). This should hold for every perturbation $\varepsilon(x)$ that satisfies the condition that the perturbed curve goes through the endpoints (x_1, y_1) and (x_2, y_2). This means that the perturbation $\varepsilon(x)$ should satisfy

$$\varepsilon(x_1) = \varepsilon(x_2) = 0. \tag{25.20}$$

Now let us perturb $y(x)$ and compute the first-order change in the length L due to that perturbation.

Problem b Use a first order Taylor expansion to show that under the perturbation $y(x) \to y(x) + \varepsilon(x)$ the integrand of expression (25.18) is to first order in ε given by:

$$\sqrt{1 + \dot{y}^2} \to \sqrt{1 + \dot{y}^2} + \frac{\dot{y}\dot{\varepsilon}}{\sqrt{1 + \dot{y}^2}}. \tag{25.21}$$

Problem c Insert this result into expression (25.18) to show that the perturbed length is to first order in $\varepsilon(x)$ given by

$$L[y+\varepsilon] = L[y] + \int_{x_1}^{x_2} \frac{\dot{y}\dot{\varepsilon}}{\sqrt{1+\dot{y}^2}}\, dx. \tag{25.22}$$

This expression states that to first order in $\varepsilon(x)$ the perturbation in the length is given by

$$\delta L = \int_{x_1}^{x_2} \frac{\dot{y}\dot{\varepsilon}}{\sqrt{1+\dot{y}^2}}\, dx. \tag{25.23}$$

The curve $y(x)$ minimizes the length when the perturbation δL in this expression vanishes for every function $\varepsilon(x)$ that satisfies the boundary conditions (25.20).

Problem d The integral in expression (25.23) can be expressed in $\varepsilon(x)$ rather than $\dot{\varepsilon}(x)$ using integration by parts. Integrate expression (25.23) by parts using the boundary conditions (25.20) to show that

$$\delta L = -\int_{x_1}^{x_2} \frac{d}{dx}\left(\frac{\dot{y}}{\sqrt{1+\dot{y}^2}}\right) \varepsilon(x)\, dx. \tag{25.24}$$

This first-order variation must vanish for every perturbation $\varepsilon(x)$ that satisfies the boundary conditions (25.20). This can be achieved only when the function that multiplies $\varepsilon(x)$ in (25.24) vanishes. The first-order variation δL therefore vanishes for every perturbation $\varepsilon(x)$ when

$$\frac{d}{dx}\left(\frac{\dot{y}}{\sqrt{1+\dot{y}^2}}\right) = 0. \tag{25.25}$$

This expression can be integrated to give $\dot{y}/\sqrt{1+\dot{y}^2} = constant$. This in turn implies that

$$\dot{y} = constant, \tag{25.26}$$

and $y(x)$ is a straight line:

$$y(x) = a + bx. \tag{25.27}$$

The constants a and b follow from the requirement that the curve goes through the endpoints (x_1, y_1) and (x_2, y_2).

Note that in this procedure the perturbation $\varepsilon(x)$ served only to describe how the length of the curve varied under a first-order perturbation of the function $y(x)$. The requirement that the first-order perturbation in expression (25.24) vanishes for all perturbations $\varepsilon(x)$ gave the differential equation (25.25) that depends on the unknown function $y(x)$ only. This procedure forms the basis of variational calculus.

25.4 The great-circle

When you live on a sphere, the shortest distance between two points is not a straight line. On an intercontinental flight between Europe and the United States, the flight path usually is curved towards the north pole. This is illustrated in Figure 25.4 which shows the path of a flight from Amsterdam to Washington DC. Amsterdam is located at about 52 degrees north, the latitude of Washington DC is about 39 degrees north, while the flight is close to Greenland at about 60 degrees north. The reason for this detour to the north is that the line with the shortest distance between two points on the sphere is not a straight line. This curve is a *great-circle*. The equator is an example of a great-circle, but on a sphere there are many other great-circles. In this section we determine the shortest curve that joins two points (φ_1, θ_1) and (φ_2, θ_2) on a sphere, where the angles φ and θ are the longitude and co-latitude on the sphere. These angles are the angles used in spherical coordinates as defined in Section 4.1.

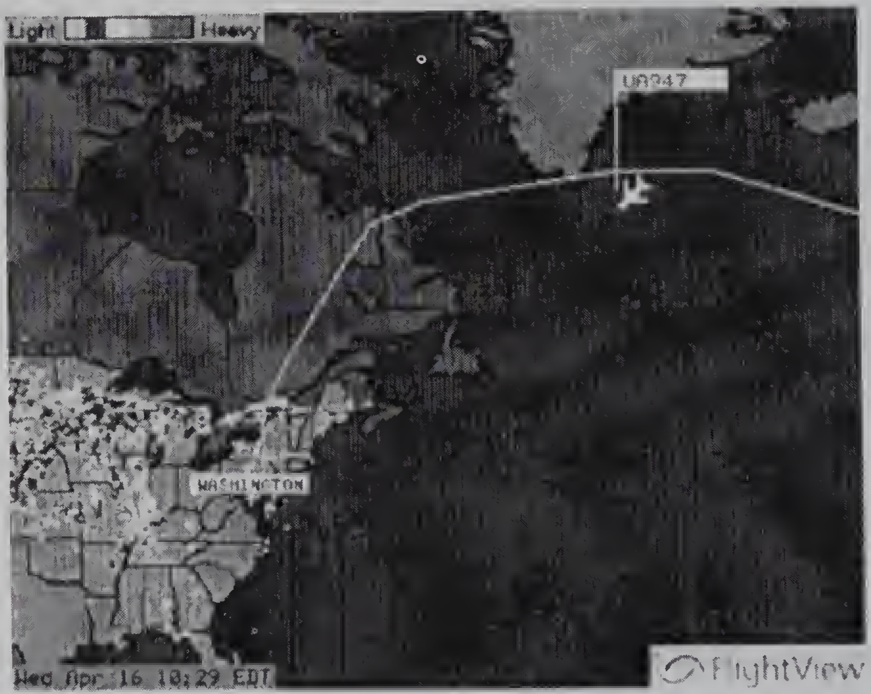

Fig. 25.4 The path of a flight from Amsterdam to Wasington DC. The planned flight path is shown in white, the true location and heading of the flight is shown by the image of the aircraft. Flight information as shown here can be viewed in real time on www.flightview.com. Figure courtesy of RLM Software.

In this section we treat the longitude φ as the independent parameter and consider the co-latitude $\theta(\varphi)$ as a function of the longitude. In order to solve this problem we first need to compute the length of the curve that joins the two endpoints for a given function $\theta(\varphi)$. For simplicity we consider the length of the curve on the unit sphere. This means that the radius r is equal to 1 and that the distance between the two points on the sphere is measured in radians.

Problem a The position vector on the unit sphere is given by the vector $\hat{\mathbf{r}}$ in expression (4.7). Show that an increment in this vector due to a change $d\varphi$ in the independent parameter is given by

$$d\hat{\mathbf{r}} = \begin{pmatrix} \cos\varphi\cos\theta\;\dot{\theta} - \sin\varphi\sin\theta \\ \sin\varphi\cos\theta\;\dot{\theta} + \cos\varphi\sin\theta \\ -\sin\theta\;\dot{\theta} \end{pmatrix} d\varphi, \tag{25.28}$$

with

$$\dot{\theta} \equiv \frac{d\theta}{d\varphi}. \tag{25.29}$$

Problem b An increment $d\varphi$ corresponds to an increments ds that satisfies $ds^2 = (d\hat{\mathbf{r}} \cdot d\hat{\mathbf{r}})$. Use this relation and expression (25.28) to show that

$$ds = \sqrt{\dot{\theta}^2 + \sin^2\theta}\; d\varphi. \tag{25.30}$$

This expression is the analogue of equation (25.16) for an increment in distance in a plane. Note that equation (25.16) depends on the derivative $\dot{y}$ only, whereas its counterpart (25.30) for a spherical geometry depends on the derivative $\dot{\theta}$ as well as on the function θ itself.

Equation (25.30) can be integrated to give the total length of the curve on the unit sphere:

$$L[\theta] = \int_{\varphi_1}^{\varphi_2} \sqrt{\dot{\theta}^2 + \sin^2\theta}\; d\varphi. \tag{25.31}$$

The length L is a functional of the function $\theta(\varphi)$. In order to find the function $\theta(\varphi)$ that minimizes the length, we perturb this function with a function $\varepsilon(\varphi)$ and require that the first-order variation of the length L with the perturbation $\varepsilon(\varphi)$ vanishes. The perturbed curve is shown by the dashed line in Figure 25.5. Since the perturbed curve must go through the endpoints (φ_1, θ_1) and (φ_2, φ_2) the perturbation must satisfy the following boundary conditions

$$\varepsilon(\varphi_1) = \varepsilon(\varphi_2) = 0. \tag{25.32}$$

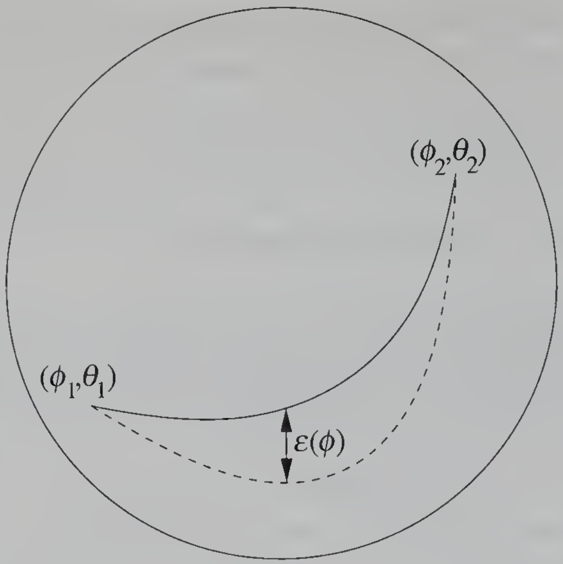

Fig. 25.5 The unperturbed curve $\theta(\varphi)$ that joins two points on the sphere (solid line), the perturbation $\varepsilon(\varphi)$, and the perturbed curve (dashed line).

Problem c Show that under the perturbation $\theta \rightarrow \theta + \varepsilon$ the following first-order perturbations hold:

$$\dot{\theta}^2 \rightarrow \dot{\theta}^2 + 2\dot{\theta}\dot{\varepsilon}, \tag{25.33}$$

$$\sin\theta \rightarrow \sin(\theta+\varepsilon) = \sin\theta + \varepsilon\cos\theta, \tag{25.34}$$

$$\sin^2\theta \rightarrow \sin^2\theta + 2\varepsilon\sin\theta\cos\theta. \tag{25.35}$$

Use these results to show that to first order in the perturbation ε the integrand of expression (25.31) is perturbed in the following way

$$\sqrt{\dot{\theta}^2 + \sin^2\theta} \rightarrow \sqrt{\dot{\theta}^2 + \sin^2\theta} + \frac{2\dot{\theta}\dot{\varepsilon} + \varepsilon\sin 2\theta}{2\sqrt{\dot{\theta}^2 + \sin^2\theta}}. \tag{25.36}$$

Integrating this expression between the endpoints φ_1 and φ_2 gives the following first-order perturbation of the length due to the perturbation $\varepsilon(\varphi)$:

$$\delta L = \frac{1}{2}\int_{\varphi_1}^{\varphi_2} \frac{2\dot{\theta}\dot{\varepsilon} + \varepsilon\sin 2\theta}{\sqrt{\dot{\theta}^2 + \sin^2\theta}}\, d\varphi. \tag{25.37}$$

Problem d This perturbation depends on both ε and $\dot{\varepsilon}$. The derivative $\dot{\varepsilon}$ can be eliminated by carrying out an integration by parts of the first term of the numerator using the boundary conditions (25.32). Use this to show that

$$\delta L = \frac{1}{2}\int_{\varphi_1}^{\varphi_2} \left\{ -\frac{d}{d\varphi}\left(\frac{2\dot{\theta}}{\sqrt{\dot{\theta}^2+\sin^2\theta}}\right) + \frac{\sin 2\theta}{\sqrt{\dot{\theta}^2+\sin^2\theta}} \right\} \varepsilon(\varphi)\, d\varphi. \quad (25.38)$$

For the curve that minimizes the distance, the length L is stationary and its first order variation δL with a perturbation $\varepsilon(\varphi)$ vanishes. This can only hold for every perturbation $\varepsilon(\varphi)$ when

$$-\frac{d}{d\varphi}\left(\frac{2\dot{\theta}}{\sqrt{\dot{\theta}^2+\sin^2\theta}}\right) + \frac{\sin 2\theta}{\sqrt{\dot{\theta}^2+\sin^2\theta}} = 0. \quad (25.39)$$

This is a differential equation for the function $\theta(\varphi)$ that gives the shortest distance between two points on a sphere. This nonlinear differential equation does not have an obvious solution. In variational calculus the resulting differential equations often are complicated, solving the differential equation can be the hardest part of the problem.

Problem e Carry out the differentiation in the first term of expression (25.39) to show that this expression can be rewritten as

$$\ddot{\theta} - 2\cot\theta\,\dot{\theta}^2 - \frac{1}{2}\sin 2\theta = 0. \quad (25.40)$$

Problem f Despite the complicated appearance of the differential equation (25.39), or the equivalent form (25.40), this equation has the following simple solution

$$\theta(\varphi) = \arctan\left(\frac{A}{\cos(\varphi+B)}\right). \quad (25.41)$$

The integration constants A and B follow from the requirement that the curve goes through the fixed endpoints, hence these constants follow from the boundary conditions $\theta(\varphi_1) = \theta_1$ and $\theta(\varphi_2) = \theta_2$. Show that the solution (25.41) indeed satisfies the differential equation (25.40). This is a lengthy calculation that requires extensive use of trigonometric identities.

The solution (25.41) gives the curve that minimizes the distance between two points on the sphere. Once the constants A and B are known, this solution can be used to compute the great-circle that goes through two fixed points. However, this function gives little insight into the shape of a great-circle. In order to understand the solution better, we rewrite the solution (25.41) as

$$\tan\theta = \frac{A}{\cos(\varphi + B)}. \tag{25.42}$$

Let us analyze this solution in Cartesian coordinates (x, y, z).

Problem g Use Figure 4.1 to show that

$$\tan\theta = \frac{\sqrt{x^2 + y^2}}{z}. \tag{25.43}$$

Problem h Show that

$$\cos(\varphi + B) = \cos\varphi\,(\cos B - \tan\varphi\sin B)\,, \tag{25.44}$$

use Figure 4.1 to show that $\cos\varphi = x/\sqrt{x^2 + y^2}$ and $\tan\varphi = y/x$, and use these results to derive that

$$\cos(\varphi + B) = \frac{x\cos B - y\sin B}{\sqrt{x^2 + y^2}}. \tag{25.45}$$

Problem i Insert these relations into expression (25.42) and show that the solution for the great-circle in Cartesian coordinates is given by

$$x\cos B - y\sin B = Az. \tag{25.46}$$

Since A and B are constant, this is the equation for a plane that goes through the origin. The points (x, y, z) on a great-circle are confined to that plane. Since the great-circle is also confined to the surface of the unit sphere, we can conclude that the great-circle is given by the intersection of a plane that goes through the center of the sphere and the endpoints of the curve, with the unit sphere. This solution is shown in Figure 25.6.

25.5 Euler–Lagrange equation

The approach taken in the previous sections was similar. The problems in both sections were reformulated into equations (25.24) and (25.38) that both give a first-order perturbation of the form $\delta L = \int (\cdots)\,\varepsilon(x)dx$. The requirement that this first-order perturbation vanishes for every perturbation $\varepsilon(x)$ led to the requirement that the terms indicated by dots must be equal to zero. The similarity of the treatments in the previous sections suggests a systematic approach.

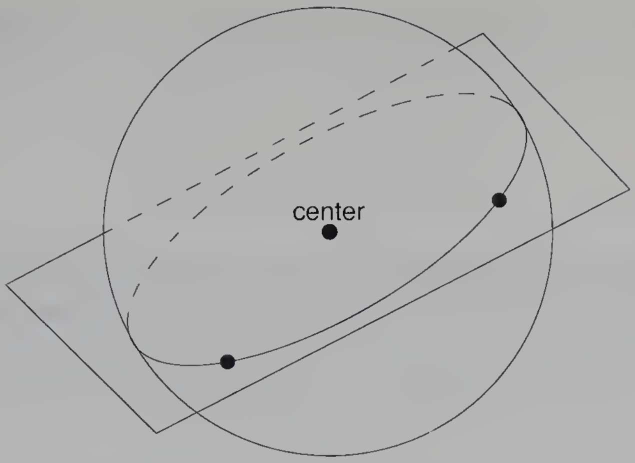

Fig. 25.6 The great-circle as the intersection of the surface of the sphere with a plane that is spanned by the center of the sphere, and the two points that define the great-circle (indicated by black circles).

Let us consider a functional L of a function $y(\xi)$:

$$L[y] = \int_{\xi_1}^{\xi_2} F(y, \dot{y}, \xi)\, d\xi. \tag{25.47}$$

In the functional (25.18) for the length of a line in a plane, the x-coordinate is the independent parameter, whereas in the functional (25.31) for the length of a curve on the sphere, φ is the independent parameter. Whatever the independent parameter is, we indicate it in this section with the variable ξ. The functional $L[y]$ depends on the function $y(\xi)$. The derivative of this function with respect to ξ is denoted by $\dot{y}$. The integrand F is assumed to depend on y, $\dot{y}$, and the independent variable ξ.

In order to find the function $y(\xi)$ that renders the functional $L[y]$ stationary, we perturb the function with a perturbation $\varepsilon(\xi)$. Under this perturbation $y \rightarrow y + \varepsilon$, and $\dot{y} \rightarrow \dot{y} + \dot{\varepsilon}$. Since $y(\xi)$ is assumed to have fixed values at the endpoints of the interval, the perturbation must vanish at these endpoints:

$$\varepsilon(\xi_1) = \varepsilon(\xi_2) = 0. \tag{25.48}$$

Problem a Use a first-order Taylor expansion to show that under this perturbation the integrand F is to first order perturbed in the following way:

$$F(y, \dot{y}, \xi) \rightarrow F(y + \varepsilon, \dot{y} + \dot{\varepsilon}, \xi) = F(y, \dot{y}, \xi) + \frac{\partial F}{\partial y}\varepsilon + \frac{\partial F}{\partial \dot{y}}\dot{\varepsilon}. \tag{25.49}$$

Note that in this Taylor expansion y and $\dot{y}$ are treated as independent parameters.

Problem b Insert this relation into expression (25.47) and derive that the first-order perturbation in the functional L is given by

$$\delta L = \int_{\xi_1}^{\xi_2} \left(\frac{\partial F}{\partial y} \varepsilon + \frac{\partial F}{\partial \dot{y}} \dot{\varepsilon} \right) d\xi . \tag{25.50}$$

Problem c Carry out an integration by parts of the second term, and show that with the boundary conditions (25.48) the first-order perturbation is given by

$$\delta L = \int_{\xi_1}^{\xi_2} \left(\frac{\partial F}{\partial y} - \frac{d}{d\xi} \left(\frac{\partial F}{\partial \dot{y}} \right) \right) \varepsilon(\xi) \, d\xi . \tag{25.51}$$

This first-order perturbation vanishes for every perturbation $\varepsilon(\xi)$ when

$$\frac{\partial F}{\partial y} - \frac{d}{d\xi} \left(\frac{\partial F}{\partial \dot{y}} \right) = 0. \tag{25.52}$$

This equation is called the *Euler–Lagrange equation*; it constitutes a differential equation for the function y that renders the functional L stationary.

The derivation of the Euler–Lagrange equation is not particularly difficult. However, when using this equation it is important to keep track of the different types of derivatives. The partial derivatives $\partial/\partial y$ and $\partial/\partial \dot{y}$ treat y and $\dot{y}$ as independent parameters. This may be confusing because $\dot{y}$ is the derivative of y with respect to ξ. The notation with partial derivatives is used to denote the dependence of F on y and $\dot{y}$, respectively. The derivative $d/d\xi$ is a total derivative. This means that the derivative acts on all quantities and the chain law must be used to compute this derivative.

Let us see how this works for the particular example of the minimization of the length of the curve in Section 25.3. According to expression (25.18) we have in that case

$$F(y, \dot{y}, x) = \sqrt{1 + \dot{y}^2}. \tag{25.53}$$

Since F does not depend explicitly on y, the partial derivative of F with respect to y vanishes: $\partial F/\partial y = 0$. Furthermore, $\partial F/\partial \dot{y} = \dot{y}/\sqrt{1 + \dot{y}^2}$. Inserting these results in the Euler–Lagrange equation and using that x plays the role of the independent variable ξ gives

$$0 - \frac{d}{dx} \left(\frac{\dot{y}}{\sqrt{1 + \dot{y}^2}} \right) = 0. \tag{25.54}$$

This equation is identical to expression (25.25). The Euler–Lagrange equation thus gives the same result as the treatment of Section 25.3.

As a second example let us consider the determination of the curve with the shortest distance between two points on a sphere. According to expression (25.31), in that case

$$F(\theta, \dot{\theta}, \varphi) = \sqrt{\dot{\theta}^2 + \sin^2\theta}. \tag{25.55}$$

In this problem, φ plays the role of the independent variable ξ, and the function y is denoted by θ. When encountering a problem with a different notation, it is often useful to reformulate the Euler–Lagrange equation for the variables that are used. In this particular problem the Euler–Lagrange equation (25.52) is given by

$$\frac{\partial F}{\partial \theta} - \frac{d}{d\varphi}\left(\frac{\partial F}{\partial \dot{\theta}}\right) = 0. \tag{25.56}$$

The partial derivatives with respect to θ and $\dot{\theta}$ are given by

$$\begin{aligned} \frac{\partial F}{\partial \theta} &= \frac{\sin\theta\cos\theta}{\sqrt{\dot{\theta}^2 + \sin^2\theta}}, \\ \frac{\partial F}{\partial \dot{\theta}} &= \frac{\dot{\theta}}{\sqrt{\dot{\theta}^2 + \sin^2\theta}}. \end{aligned} \tag{25.57}$$

Inserting these values in the Euler–Lagrange equation (25.56) gives

$$\frac{\sin\theta\cos\theta}{\sqrt{\dot{\theta}^2 + \sin^2\theta}} - \frac{d}{d\varphi}\left(\frac{\dot{\theta}}{\sqrt{\dot{\theta}^2 + \sin^2\theta}}\right) = 0. \tag{25.58}$$

Problem d Verify that this expression is identical to equation (25.39).

The examples shown here imply that the Euler–Lagrange equation gives the same results as we derived in the previous sections. We have covered the simplest case where $y(\xi)$ depends on one variable only and where the endpoints of the function y are fixed. The treatment can be generalized to include functions y that depend on more than one independent variable, functions whose endpoints are not necessarily fixed, and functionals that also contain the second derivative $\ddot{y}$. Details can be found in reference [95].

The treatment of this section can be generalized for the situation that the functional depends on several different variables. This is for example the case when one wants to minimize a functional that depends on the trajectory $\mathbf{r}(\xi)$ that depends on three spatial coordinates. In that case, the functional depends on the coordinates $\mathbf{r}$:

$$F = F[\mathbf{r}], \tag{25.59}$$

where $\mathbf{r}$ is a function of the independent parameter ξ. The curve that renders this functional stationary can be found by changing the trajectory $\mathbf{r}(\xi)$ with a vector perturbation $\boldsymbol{\varepsilon}(\xi)$, and by requiring that to first order in $\boldsymbol{\varepsilon}(\xi)$ the functional does not change. This perturbation in general has three components as well. However, nothing keeps us from perturbing one of the components of $\boldsymbol{\varepsilon}$ only, let us denote this component with ε_i. Let the components of the vector $\mathbf{r}$ be denoted by y_i. The derivation in expressions (25.48)–(25.52) can be repeated for that case, with the only change that the variable ε must be replaced by ε_i. This gives the Euler–Lagrange equation for each of the components y_i:

$$\frac{\partial F}{\partial y_i} - \frac{d}{d\xi}\left(\frac{\partial F}{\partial \dot{y}_i}\right) = 0. \tag{25.60}$$

25.6 Lagrangian formulation of classical mechanics

In this section we consider a particle with mass m that moves in one dimension. The mass m is not necessarily constant. In order to relate this problem to the notation of the previous section we denote the location of the particle with the coordinate y. The particle starts at a location y_1 at time t_1 and moves to the endpoint y_2 where it arrives at time t_2. The kinetic energy of the particle is given by $E_K = m\dot{y}^2/2$, where the dot denotes the time derivative: $\dot{y} \equiv dy/dt$. The time t now plays the role of the independent variable ξ. The potential energy of the particle is given by $E_P = V(y)$, where V describes the potential in which the particle moves. Let us consider the difference of the kinetic and potential energy integrated over time:

$$A[y] \equiv \int_{t_1}^{t_2} \left(\frac{1}{2}m\dot{y}^2 - V(y)\right) dt, \tag{25.61}$$

this quantity is called the *action* [44]. At this moment there is no specific reason yet to compute the action, but we will see that the trajectory $y(t)$ that renders the action stationary has a special meaning.

In the notation of equation (25.47) the integrand of the action is given by

$$L(y, \dot{y}, t) = \frac{1}{2}m\dot{y}^2 - V(y), \tag{25.62}$$

this quantity is called the *Lagrangian* [44].

Problem a Verify that the partial derivatives of the Lagrangian with respect to y and $\dot{y}$ are given by

$$\begin{aligned} \frac{\partial L}{\partial y} &= -\frac{\partial V}{\partial y}, \\ \frac{\partial L}{\partial \dot{y}} &= m\dot{y}. \end{aligned} \tag{25.63}$$

Problem b Insert these results in the Euler–Lagrange equation to show that the trajectory $y(t)$ that renders the action stationary is given by

$$\frac{d}{dt}(m\dot{y}) = -\frac{\partial V}{\partial y}. \tag{25.64}$$

The derivative $-\partial V/\partial y$ is simply the force F that acts on the particle. Since $\dot{y}$ is the velocity, $m\dot{y}$ is the momentum, and equation (25.64) is equivalent to Newton's law: $d(mv)/dt = F$. This result gives us a new way to interpret the motion of a particle in classical mechanics. The Lagrangian (25.62) is the difference of the kinetic and potential energy. This means that in classical mechanics a particle follows a trajectory such that the time-averaged difference between the kinetic and potential energy is minimized. This is one example of the variational formulation of classical mechanics. More details and many examples can be found in references [44, 60].

Let us redo this problem in three dimensions. In that case the action is given by

$$A \equiv \int_{t_1}^{t_2} \left(\frac{1}{2} m\dot{\mathbf{y}}^2 - V(\mathbf{y}) \right) dt, \tag{25.65}$$

where $\mathbf{y}(t)$ describes the trajectory of the particle as a function of time.

Problem c Use the Euler–Lagrange equation (25.60) to show that this functional is stationary when

$$\frac{d}{dt}(m\dot{y}_i) = -\frac{\partial V}{\partial y_i}. \tag{25.66}$$

In vector-form this expression can be written as

$$\frac{d}{dt}(m\dot{\mathbf{y}}) = -\nabla V. \tag{25.67}$$

This is Newton's law in three dimensions for a force that is related to the potential by $\mathbf{F} = -\nabla V$.

25.7 Rays are curves of stationary travel time

In Section 12.4 we discussed geometric ray theory as a high-frequency approximation to solutions of the wave equation. In this section we develop an alternative view on ray theory. When a wave travels along a segment of a ray with length ds, and if the wave velocity is denoted by c, then the time needed to cover this distance is given by $dt = c^{-1} ds$. In this section we use the symbol u for slowness, the reciprocal of velocity:

$$u \equiv 1/c. \tag{25.68}$$

Using this definition, the total travel time along the ray is given by

$$T = \int u(\mathbf{r})\, ds. \tag{25.69}$$

In this section we consider the trajectories that render the travel time stationary.

The integral (25.69) suggests use of the arc-length s as independent parameter. This is, however, not a good idea since it is not known how long the ray that connects a source to a receiver will be. For this reason the endpoint condition (25.48) cannot be applied because it is not known what ξ_2 is. The travel time cannot be used as independent parameter either, because it is the travel time that we seek to optimize. Instead we will use the relative arc-length as independent parameter. Let the total length of the ray be denoted by S. The relative arc-length is defined as

$$\xi \equiv s/S. \tag{25.70}$$

At the starting point of the ray $\xi = 0$, while at the endpoint $\xi = 1$.

At this moment we do not know S yet, but this quantity can be related to the curve $\mathbf{r}(\xi)$ by using that

$$ds = \left| \frac{d\mathbf{r}}{d\xi} \right| d\xi = |\dot{\mathbf{r}}|\, d\xi = \sqrt{\dot{x}^2 + \dot{y}^2 + \dot{z}^2}\, d\xi. \tag{25.71}$$

Problem a Use these results to show that the travel time can be written as

$$T[\mathbf{r}] = \int_0^1 F(\mathbf{r}, \dot{\mathbf{r}})\, d\xi, \tag{25.72}$$

with

$$F(\mathbf{r}, \dot{\mathbf{r}}) = u(\mathbf{r})\sqrt{\dot{x}^2 + \dot{y}^2 + \dot{z}^2}. \tag{25.73}$$

Problem b The curve that renders the travel time stationary follows from the Euler–Lagrange equation (25.60). Show that

$$\frac{\partial F}{\partial x} = \frac{\partial u}{\partial x}\sqrt{\dot{x}^2 + \dot{y}^2 + \dot{z}^2}, \tag{25.74}$$

and

$$\frac{\partial F}{\partial \dot{x}} = u\frac{\dot{x}}{\sqrt{\dot{x}^2 + \dot{y}^2 + \dot{z}^2}}. \tag{25.75}$$

Problem c Use these results to show that the Euler–Lagrange equation (25.60) for the variable x is given by

$$\frac{d}{d\xi}\left(u\frac{\dot{x}}{\sqrt{\dot{x}^2 + \dot{y}^2 + \dot{z}^2}}\right) = \frac{\partial u}{\partial x}\sqrt{\dot{x}^2 + \dot{y}^2 + \dot{z}^2}. \tag{25.76}$$

Similar equations hold for the variables y and z. In order to interpret this equation, we introduce the unit vector $\hat{\mathbf{n}}$, whose x-component is defined by

$$n_x \equiv \frac{\dot{x}}{\sqrt{\dot{x}^2 + \dot{y}^2 + \dot{z}^2}}, \tag{25.77}$$

with a similar definition for the y- and z-components.

Problem d Show that this vector is of unit length: $|\hat{\mathbf{n}}| = 1$.

Problem e Expression (25.71) can be used to convert the ξ-derivative in expression (25.76) into a derivative along the curve, by using

$$\frac{d}{d\xi} = \frac{ds}{d\xi}\frac{d}{ds} = \sqrt{\dot{x}^2 + \dot{y}^2 + \dot{z}^2}\,\frac{d}{ds}. \tag{25.78}$$

Use these results to show that the Euler–Lagrange equation (25.76) can be written as

$$\frac{d}{ds}(un_x) = \frac{\partial u}{\partial x}. \tag{25.79}$$

Similar expressions hold for the y- and z-components of the Euler–Lagrange equation. In vector-form the Euler–Lagrange equation can be written as

$$\frac{d}{ds}(u\hat{\mathbf{n}}) = \nabla u. \tag{25.80}$$

With definition (25.77), the unit vector **n** can be rewritten in the following way

$$\hat{\mathbf{n}} = \frac{d\mathbf{r}/d\xi}{\sqrt{\dot{x}^2 + \dot{y}^2 + \dot{z}^2}} = \frac{d\mathbf{r}}{d\xi}\frac{d\xi}{ds} = \frac{d\mathbf{r}}{ds}, \tag{25.81}$$

where expression (25.78) is used in the second identity. Using this result in equation (25.80), and using that $u = c^{-1}$ gives

$$\frac{d}{ds}\left(\frac{1}{c}\frac{d\mathbf{r}}{ds}\right) = \nabla\left(\frac{1}{c}\right). \tag{25.82}$$

This equation is called the *equation of kinematic ray tracing.*

The equation of kinematic ray tracing can be derived from the eikonal equation (12.25) that was derived from the high-frequency approximation to the wave equation [3]. The eikonal equation (12.25) describes the propagation of wave-fronts, while the equation of kinematic ray tracing (25.82) describes the propagation of rays. These equations are equivalent and provide a complementary view of geometric ray theory. In this section we derived the equation of kinematic ray tracing from the requirement that the travel time is stationary. This means that rays are curves that render the travel time stationary. Most often, rays are curves that minimize the travel time. However, seismic waves that reflect once off the Earth's surface are so-called *minimax arrivals*. The travel time increases when the ray is perturbed in one direction and it decreases when the ray is perturbed in another direction. Whether a ray is a minimum time arrival or a minimax arrival can be detected in the phase of the arriving waves. Observations of this phenomenon are shown by Choy and Richards [25].

It is interesting to convert the derivative d/ds in expression (25.82) into a time derivative by using that

$$\frac{d}{ds} = \frac{dt}{ds}\frac{d}{dt} = \frac{1}{c}\frac{d}{dt}. \tag{25.83}$$

This changes the equation of kinematic ray tracing into

$$\frac{d}{dt}\left(\frac{1}{c^2}\frac{d\mathbf{r}}{dt}\right) = c\nabla\left(\frac{1}{c}\right). \tag{25.84}$$

This expression can be compared with Newton's law (25.67) for a particle with a mass m that may vary with time:

$$\frac{d}{dt}\left(m\frac{d\mathbf{r}}{dt}\right) = \mathbf{F}. \tag{25.85}$$

These equations are identical when $1/c^2$ in (25.84) is equated to the mass m in (25.85), and when $c\nabla\left(1/c\right)$ takes the role of the force **F** in Newton's law. In classical mechanics it is the force that makes a trajectory curve. Therefore, in geometric ray theory it is the velocity gradient that makes a ray curve. The close analogy between geometric ray theory and classical mechanics has led to a wide body of theory that utilizes the same mathematical tools for classical mechanics and geometric ray theory [11, 38].

25.8 Lagrange multipliers

Often in minimization there is a constraint that the solution should satisfy. In Section 25.10 we consider the problem of a wire that is suspended between two points. The shape of the wire is dictated by the requirement that its potential energy in the gravitational field is minimized. Obviously, the potential energy is smallest when the wire curves down as much as possible. The constraint that the wire has a given length restricts the downward displacement of the wire. This amounts to a *constrained optimization problem* where the potential energy is minimized under the constraints that the wire has a fixed length. Problems like this can be solved with a technique called *Lagrange multipliers*. In this section we introduce this technique first with a finite-dimensional problem, and then generalize it to problems with infinitely many degrees of freedom.

To fix our mind we consider the following question: what is the point on the surface of a ball that has the lowest potential energy in a gravitational field? This is a trivial problem, and the solution obviously is given by the point at the bottom of the ball. Here we solve this problem while introducing the technique of Lagrange multipliers.

Let us first cast the condition that the point lies on the surface of the ball in a mathematical form. When the ball has a radius R and is centered on the origin, this constraint is given by

$$C(x, y, z) = x^2 + y^2 + z^2 = R^2. \tag{25.86}$$

The constraint thus implies that $C(x, y, z) = constant$. Suppose that we move a point $\mathbf{r} = (x, y, z)$ along the surface of the ball with an infinitesimal displacement $\delta\mathbf{r}$, as shown in Figure 25.7. Since the displaced point $\mathbf{r} + \delta\mathbf{r}$ must lie on the surface of the ball as well, it must satisfy

$$C(\mathbf{r} + \delta\mathbf{r}) = C(\mathbf{r}) \tag{25.87}$$

Problem a Use equation (5.8) to show that this implies that

$$\nabla C \cdot \delta\mathbf{r} = 0. \tag{25.88}$$

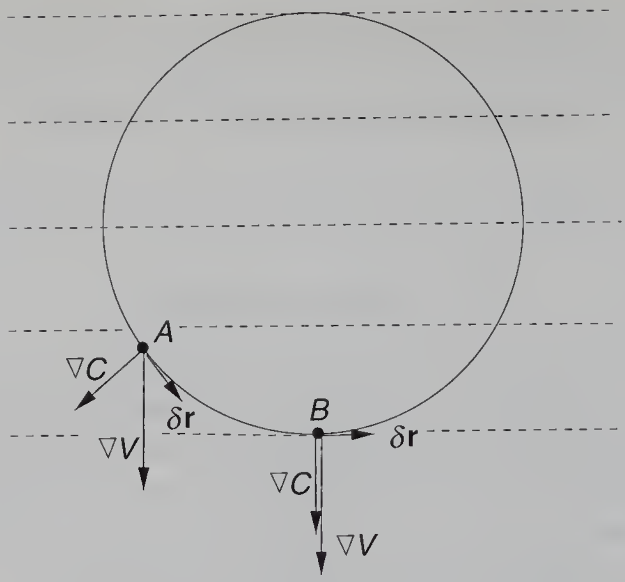

Fig. 25.7 The direction of the vectors ∇C and ∇V at two points on a ball. The dashed lines indicate the equipotential surfaces where $V = constant$.

Geometrically this condition states that $\delta \mathbf{r}$ is perpendicular to the gradient ∇C. For our particular problem, ∇C points in the radial direction, so the condition (25.87) simply states that the point is perturbed along the surface of the sphere.

We want to minimize the potential energy $V(\mathbf{r})$ associated with the gravitational field as a function of the location $\mathbf{r}$. Under a perturbation $\delta \mathbf{r}$, the potential energy is to first order changed by an amount

$$\delta V = (\nabla V \cdot \delta \mathbf{r}). \tag{25.89}$$

The potential energy is minimized when this quantity vanishes, this is the case when the displacement $\delta \mathbf{r}$ is perpendicular to ∇V.

How can we see that the point A in the sphere in Figure 25.7 does not minimize the potential energy, but that the point B does minimize the potential energy? Consider the displacement $\delta \mathbf{r}$ at point A along the surface of the sphere. At point A the displacement $\delta \mathbf{r}$ is not perpendicular to the gradient ∇V. According to expression (25.89) this implies that the potential energy changes to first order under this perturbation, hence point A does not minimize the potential energy. At point B, the displacement $\delta \mathbf{r}$ is perpendicular to the vector ∇V, so that at that point the potential energy does not change to first order under this perturbation. In other words, point B is an extremum of the potential energy.

At point A the vectors ∇C and ∇V are not parallel; the vector ∇C points in the radial direction while the vector ∇V points downward. However, at the minimum of the potential energy at point B, these vectors are parallel, see Figure 25.7. This means that at the minimum the gradient vectors satisfy

$$\nabla C = -\lambda \nabla V. \tag{25.90}$$

The constant $-\lambda$ accounts for the fact that the vectors in general have a different length. The minus sign is introduced in order to conform to the notation most often found in the literature; it has no specific meaning. The constant λ is called the *Lagrange multiplier*. Equation (25.90) can also be written as

$$\nabla (V + \lambda C) = 0. \tag{25.91}$$

For the unconstrained minimization of the potential energy, we just needed to solve the equation $\nabla V = 0$. The constraint that the point lies on the surface of the ball leads to the minimization problem (25.91). Note that this problem contains the unknown parameter λ, hence the constrained minimization problem depends on the four variables x, y, z, and λ, whereas the original problem was dependent only on the three variables x, y, and z. However, the constrained minimization problem has an extra equation that should be satisfied: the constraint (25.86). This means that we now have four equations with four unknowns.

Yet another way to see that point B minimizes the potential energy is the following. The constraint states that the point lies on the surface of the sphere. Stated differently, the point must lie on the surface $C(x, y, z) = constant$. At the extremum of the potential energy, this surface must touch the surface $V(x, y, z) = constant$; if it would intersect that surface at an angle the point could be moved along the surface $C(x, y, z) = constant$ while raising or lowering the potential energy. This contradicts the fact that the point maximizes or minimizes the potential energy. This means the extremum of the potential energy is attained at the points where the surfaces $C(x, y, z) = constant$ and $V(x, y, z) = constant$ are tangent. (When a ball lies on the ground, the surface of the ball and the ground touch each other.) A different way of stating this is that the vectors ∇C and ∇V are parallel, which is equivalent to condition (25.90).

Let us see how the machinery of the Lagrange multipliers works for finding the point on a ball that has the lowest potential energy. In this case the potential energy is given by

$$V(\mathbf{r}) = z, \tag{25.92}$$

and the constraint that the point lies on a sphere with radius R is given by (25.86). For this example, $V + \lambda C = z + \lambda(x^2 + y^2 + z^2)$.

Problem b Show that for this problem equation (25.91) leads to

$$\begin{pmatrix} 2\lambda x \\ 2\lambda y \\ 1 + 2\lambda z \end{pmatrix} = 0. \tag{25.93}$$

This expression constitutes three equations for the four unknowns x, y, z, and λ. The first two lines dictate that $x = y = 0$. The last line states that $z = -1/2\lambda$.

Problem c Use the constraint (25.86) to show that $\lambda = \pm 1/2R$.

This means that the solution is given by

$$\mathbf{r} = \begin{pmatrix} 0 \\ 0 \\ \pm R \end{pmatrix}. \tag{25.94}$$

The potential energy has extrema for the top and the bottom of the sphere. The top of the sphere maximizes the potential energy, while at the bottom of the sphere the potential energy has a minimum.

In the treatment of this section, it is not essential that the constraint $C(\mathbf{r})$ describes the surface of the sphere, and the $V(\mathbf{r})$ is the potential energy. The employed arguments hold for the minimization of any function $V(\mathbf{r})$ under a constraint $C(\mathbf{r})$. As long as the gradient of these functions is well-defined the constrained optimization problem can be solved with expression (25.91).

The extension of Lagrange multipliers to the constrained optimization of functionals is not trivial. Here we sketch the main idea. Suppose we want to minimize the functional

$$L[y] = \int F(y, \dot{y}, \xi)\, d\xi, \tag{25.95}$$

subject to the constraint

$$\int C(y, \dot{y}, \xi) d\xi = constant. \tag{25.96}$$

Suppose that $y(\xi)$ is modified with a perturbation $\varepsilon(\xi)$ such that the perturbed function also satisfies the constraint (25.96). This means that the perturbation of the constraint is given by

$$\int \left(\frac{\partial C}{\partial y}\varepsilon + \frac{\partial C}{\partial \dot{y}}\dot{\varepsilon} \right) d\xi = 0. \tag{25.97}$$

Using an integration by parts, as in Section 25.5, then gives

$$\int \left(\frac{\partial C}{\partial y} - \frac{d}{d\xi}\left(\frac{\partial C}{\partial \dot{y}} \right) \right) \varepsilon(\xi)\, d\xi = 0. \tag{25.98}$$

This equation is equivalent to expression (25.87) for a finite-dimensional problem. With a similar argument, the minimization of the functional (25.95) leads to the condition

$$\int \left(\frac{\partial F}{\partial y} - \frac{d}{d\xi} \left(\frac{\partial F}{\partial \dot{y}} \right) \right) \varepsilon(\xi)\, d\xi = 0. \tag{25.99}$$

The last two expressions state that the functions $\left(\partial C/\partial y - \dfrac{d}{d\xi} (\partial C/\partial \dot{y}) \right)$ and $\left(\partial F/\partial y - \dfrac{d}{d\xi} (\partial F/\partial \dot{y}) \right)$ are both "perpendicular" to the perturbation $\varepsilon(\xi)$, just like ∇C and ∇V were perpendicular to the perturbation $\delta \mathbf{r}$ of the point on the sphere. Since this must hold for any perturbation $\varepsilon(\xi)$, this means that these functions are "parallel" to each other:

$$\left(\frac{\partial F}{\partial y} - \frac{d}{d\xi} \left(\frac{\partial F}{\partial \dot{y}} \right) \right) = -\lambda \left(\frac{\partial C}{\partial y} - \frac{d}{d\xi} \left(\frac{\partial C}{\partial \dot{y}} \right) \right). \tag{25.100}$$

The is the analogue of equation (25.90). This argument is not rigorous, the quotation marks highlight a vague use of the concepts perpendicular and parallel. A more thorough (and more complex) derivation is given by Smith [95]. However, accepting expression (25.100), the constrained optimization problem is given by the equation

$$\frac{\partial (F + \lambda C)}{\partial y} - \frac{d}{d\xi} \left(\frac{\partial (F + \lambda C)}{\partial \dot{y}} \right) = 0. \tag{25.101}$$

This expression is nothing but the Euler–Lagrange equation for the function $F + \lambda C$. The constrained optimization problem can thus be solved by optimizing $F + \lambda C$. This introduces the additional parameter λ. This parameter can be found by solving the constraint (25.96).

25.9 Designing a can with an optimal shape

In this section we return to the problem of Sections 25.1 and 25.2 where we treated the design of the shape of a can that maximizes the content for a given surface area. We treat this problem here using the Lagrange multiplier. Let us consider the design problem in a slightly different way by assuming that the surface area S of the bottom and the top of the can is fixed. We noted in Section 25.2 that the volume of the can is largest when the circumference C is minimized for a fixed value of the surface S of the top and bottom.

The circumference is related to the radius $r(\varphi)$ by expression (25.14). The surface area of the top and bottom is given by

$$S = \frac{1}{2}\int_0^{2\pi} r^2 d\varphi. \tag{25.102}$$

According to the theory of the previous section, this constrained optimization problem can be solved by optimizing $C + \lambda S$, where the Lagrange multiplier λ follows from the constraint (25.102).

Problem a Use expressions (25.14) and (25.102) to show that the constrained optimization problem is solved by solving the Euler–Lagrange equation for

$$F(r, \dot{r}) = \sqrt{\dot{r}^2 + r^2} + \frac{\lambda}{2} r^2. \tag{25.103}$$

Problem b Show that

$$\frac{\partial F}{\partial r} = \frac{r}{\sqrt{\dot{r}^2 + r^2}} + \lambda r, \tag{25.104}$$

and

$$\frac{\partial F}{\partial \dot{r}} = \frac{\dot{r}}{\sqrt{\dot{r}^2 + r^2}}. \tag{25.105}$$

Problem c Take the total derivative of the previous expression with respect to φ and show that the Euler–Lagrange equation for this optimization problem is given by

$$\frac{\ddot{r}r^2 - \dot{r}^2 r}{\left(\dot{r}^2 + r^2\right)^{3/2}} = \frac{r}{\sqrt{\dot{r}^2 + r^2}} + \lambda r. \tag{25.106}$$

Problem d This is a nonlinear differential equation. However, we know from Section 25.2 that a constant radius is likely to be a solution. Show that the solution $r(\varphi) = \mathit{constant}$ reduces equation (25.106) to

$$1 + \lambda r = 0. \tag{25.107}$$

Problem e This condition implies that $r = -\lambda^{-1}$. Determine the Lagrange multiplier λ by inserting this expression into the constraint (25.102) to show that

$$\lambda = \sqrt{\pi/S}, \tag{25.108}$$

and that this finally gives the solution

$$r = \sqrt{S/\pi}. \tag{25.109}$$

This example shows that the theory of the Lagrange multiplier states that a cylindrical can has the largest volume for a given surface area.

25.10 Chain line and the Gateway Arch of St. Louis

In this section we apply the theory of Lagrange multipliers to the problem of finding the shape of a wire of length L that is suspended between the points $(0, 0)$ and $(X, 0)$, as shown in Figure 25.8. Following equation (25.18), the total length of the wire is given by

$$L = \int_0^X \sqrt{1+\dot{y}^2}\, dx. \tag{25.110}$$

This expression formulates the constraint that the wire has a given length L. In this section we use x as the independent variable and $y(x)$ as the dependent variable. The variable ρ denotes the mass of the wire per unit length. An increment dx corresponds to an increment in length given by $\sqrt{1+\dot{y}^2}\, dx$, the corresponding mass is $\rho\sqrt{1+\dot{y}^2}\, dx$, and the associated potential energy is given by $g\rho y\sqrt{1+\dot{y}^2}\, dx$, with g the acceleration of gravity. The potential energy of the wire thus is given by

$$V[y] = g\rho \int_0^X y\sqrt{1+\dot{y}^2}\, dx. \tag{25.111}$$

If the wire has no stiffness, it will assume a shape that minimizes its potential energy. The shape of the wire can thus be found by minimizing the potential energy (25.111) under the constraint (25.110) that its length is fixed. The curve $y(x)$ that minimizes $\int_0^X y\sqrt{1+\dot{y}^2}\, dx$ also minimizes $g\rho \int_0^X y\sqrt{1+\dot{y}^2}\, dx$. For this reason the constants $g\rho$ are left out in the following part of this section.

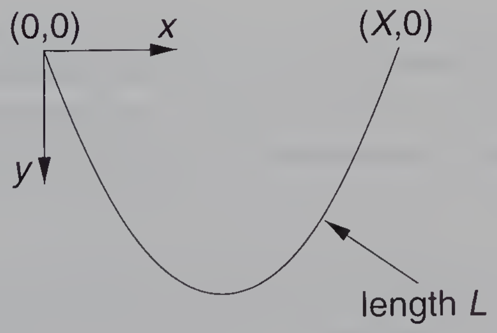

Fig. 25.8 The geometry of a wire suspended between two points.

According to the discussion of Section 25.8 the solution of this constrained optimization problem can be found by optimizing $V + \lambda L$, and by using the constraint (25.110) to find the Lagrange multiplier λ.

Problem a Use the theory of the previous section to show that the shape of the wire is found by optimizing

$$F[y] = \int_0^X \mathcal{F}(y, \dot{y})\, dx, \tag{25.112}$$

with

$$\mathcal{F}(y, \dot{y}) = y\sqrt{1 + \dot{y}^2} + \lambda\sqrt{1 + \dot{y}^2} \tag{25.113}$$

Problem b Show that

$$\frac{\partial \mathcal{F}}{\partial y} = \sqrt{1 + \dot{y}^2}, \tag{25.114}$$

and

$$\frac{\partial \mathcal{F}}{\partial \dot{y}} = \frac{\dot{y}}{\sqrt{1 + \dot{y}^2}}(y + \lambda). \tag{25.115}$$

Problem c Take the total derivative of the last expression to derive that

$$\frac{d}{dx}\left(\frac{\partial \mathcal{F}}{\partial \dot{y}}\right) = \frac{1}{\left(1 + \dot{y}^2\right)^{3/2}} \left\{\ddot{y}(y + \lambda) + \dot{y}^2\sqrt{1 + \dot{y}^2}\right\}. \tag{25.116}$$

Remember that in this derivative all variables that depend on x are differentiated.

Problem d Show that the Euler–Lagrange equation for this problem can be written as

$$1 + \dot{y}^2 = \ddot{y}(y + \lambda). \tag{25.117}$$

This is a non-linear differential equation that displays the complexity that often results in the solution of variational problems. There is no simple way to guess the solution, or to construct it in a systematic way. For this reason we simply state the following trial solution

$$y(x) = A\cosh\left(B(x + h)\right) + C, \tag{25.118}$$

where A, B, C, and h are constants.

Problem e Insert this solution into the differential equation (25.117) and show that the result can be written as

$$(1 - A^2B^2) = AB^2(\lambda + C)\cosh(B(x + h))\,. \tag{25.119}$$

The left-hand side of this expression is independent of x, while the right-hand side varies with x. This is possible only when the coefficients of the left-hand side and the right-hand side vanish. This implies that

$$1 - A^2B^2 = 0 \qquad \text{and} \qquad AB^2(\lambda + C) = 0. \tag{25.120}$$

These equations are satisfied when $A = 1/B$ and $C = -\lambda$, so that the solution is given by

$$y(x) = \frac{1}{B}\cosh(B(x + h)) - \lambda. \tag{25.121}$$

This expression contains three constants. There are, however, three pieces of information that we have not used yet: the boundary conditions which state the wire goes through the points $(0, 0)$ and $(X, 0)$, and the constraint (25.110) which states that the wire has length L.

Problem f Use the boundary conditions to show that $h = -X/2$ and $\lambda = B^{-1}\cosh(X/2)$, so that the solution is given by

$$y(x) = B^{-1}\left\{\cosh\left(B\left(x - \frac{X}{2}\right)\right) - \cosh\left(\frac{BX}{2}\right)\right\}. \tag{25.122}$$

Problem g Insert this solution into the constraint (25.110) and show that this leads to the following equation for the constant B:

$$\frac{2}{B}\sinh\left(\frac{BX}{2}\right) = L \tag{25.123}$$

This is a transcendental equation that cannot be solved in closed form, but it can be solved numerically. This completes the solution of finding the shape of a wire suspended between two points.

It follows from equation (25.122) that the wire is shaped like the cosine-hyperbolic. This curve is called the *catenary*; this word is derived from the Latin word *catena* that means chain. The catenary can be seen in the chain that is suspended in the background of Figure 25.9. Knowing the shape of a suspended wire is more than a mathematical curiosity. A suspended wire has in general such a

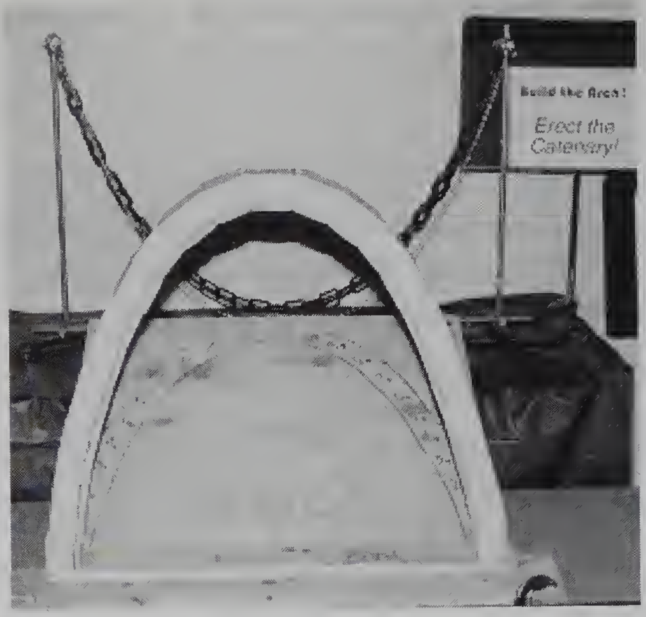

Fig. 25.9 The chain in the background, and the arch in the foreground have the shape of a catenary. The arch consists of blocks of foam that are not glued together. The figure is made by high-school students and is reproduced with permission of the New Trier Connections Project.

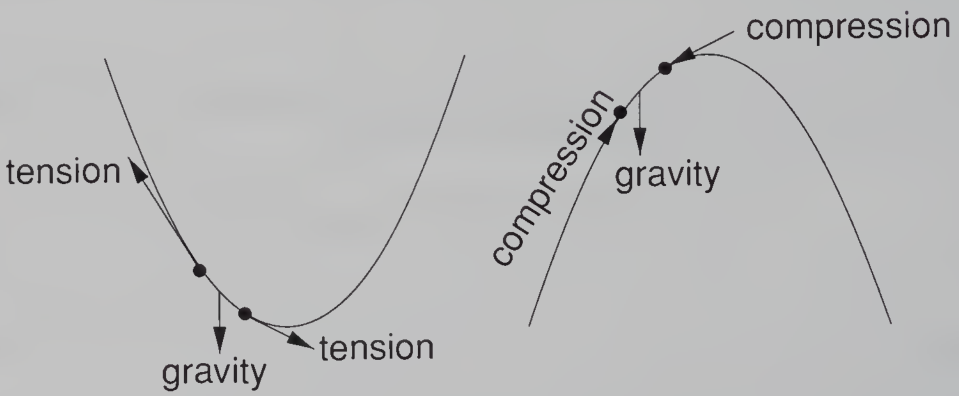

Fig. 25.10 The balance of the tensional forces and gravity for a suspended wire (left panel), and the balance between tensional forces and the gravitational force for an arch (right panel).

small stiffness that this property can be ignored. (In the analysis of this section the wire does not have any stiffness.) This means that the wire does not support any internal bending moments. The shape of the wire is dictated by the balance of the gravitational force and the net tensional force that arises from the curvature of the wire as shown in the left panel of Figure 25.10.

Let us suppose we turn the wire upside down and that we build an arch. In that case the gravitational force acts into the arch and the tensional forces of the left panel of Figure 25.10 are changed into compressional forces that balances the gravitational force, as shown in the right panel of Figure 25.10. This implies that an arch built in the shape of a catenary does not require any internal bending moments to maintain its shape. This is illustrated with the arch shown in Figure 25.9 that was built by students of the New Trier High School in Winnetka, Illinois. This arch consists of pieces of foam that are cut in a form to jointly give the shape of a catenary. The pieces of foam are not glued together. Since a catenary arch does not need any bending moment to support itself, there is no shear force that makes the blocks slide away from each other; all internal force acts along the curve of the arch itself. In this sense the arch is truly self-supporting. The Gateway Arch in St. Louis, with a height of 210 m, is built in the shape of a catenary. This allows for the slender design of this monument.

26

Epilogue, on power and knowledge

We all continue to feel a frustration because of our inability to foresee the soul's ultimate fate. Although we do not speak about it, we all know that the objectives of our science are, from a general point of view, much more modest than the objectives of, say, the Greek sciences were; that our science is more successful in giving us power than in giving us knowledge of truly human interest. [E. P. Wigner, 1972, The place of consciousness in modern physics, in *Consciousness and Reality*, eds. C. Muses and A. M. Young, New York, Outerbridge and Lizard, pp. 132–141].

In this book we have explored many methods of mathematics as used in the physical sciences. Mathematics plays a central role in the physical sciences because it is the only language we have for expressing quantitative relations in the world around us. In fact, mathematics not only allows us to express phenomena in a quantitative way, it also has a remarkable predictive power in the sense that it allows us to deduce the consequences of natural laws in terms of measurable quantities. In fact, we do not quite understand *why* mathematics gives such an accurate description of the world around us [120].

It is truly stunning how accurate some of the predictions in (mathematical) physics have been. The orbits of the planetary bodies can now be computed with an extreme accuracy. Morrison and Stephenson [72] compared the path of a solar eclipse at 181 BC with historic descriptions made in a city in eastern China that was located in the path of the solar eclipse. According to the computations, the path of the solar eclipse passed 50 degrees west of the site of this historic observation. This eclipse took place about 2000 years ago; this means that the Earth has rotated through about 2.8×10^8 degrees since the eclipse. The *relative* error in the path of the eclipse over the Earth is thus only 1.7×10^{-7}. In fact, this discrepancy of 50 degrees can be explained well by the observed deceleration of the Earth [72] due to the braking effect of the Earth's tides.

The light emitted by hydrogen atoms has discrete spectral lines which are due to the fact that electrons behave as standing waves. Every electron is coupled to the field of electromagnetic radiation (light). There is a small chance that an electron will emit and re-absorb a virtual photon [91]. (Photons are the light-quanta.) This leads to the so-called *Lamb shift* of the spectral lines of light radiated by excited hydrogen atoms. For hydrogen atoms the shift for the transition between the 2s and $2p_{1/2}$ state is 1060 MHz. This corresponds to a reciprocal wavelength of 0.035 cm^{-1}, the reciprocal wavelength that corresponds to the ionization energy is 2700 cm^{-1}. Compared to the ionization energy of the ground state this corresponds therefore to a *relative* frequency shift of 1.3×10^{-6}. This is in very good agreement with observations. It is interesting to note that this prediction of the Lamb shift is based on second order perturbation theory [91]. This means that an approximate theory provides a stunningly accurate prediction of the Lamb shift.

A third example of the extreme accuracy of mathematics in the physical sciences is the perihelion precession of Mercury [76]. According to the laws of Newton, a planet will orbit in a fixed ellipsoidal orbit around the Sun. The general theory of relativity predicts that this ellipse slowly changes its position; the point of the ellipse closest to the Sun (the perihelion) slowly precesses around the Sun. According to the theory of general relativity this precession is given by 42.98 arcsec/century, whereas the observed precession rate is 43.1 ± 0.1 arcsec/century. Note that this precession rate is extremely small, but that it is well predicted from theory.

Mathematics not only provides us with valuable and stunningly accurate insights in the world around us, it is also an indispensable tool in making technical innovations. The design and implementation of rockets, aircraft, chemical plants, water treatment systems, modern electronics, information technology, and many other innovations would have been impossible without mathematics. Mathematics and the physical sciences have created many new opportunities for mankind. For this reason one can state that mathematics and the physical sciences have greatly increased our power to modify the world in which we live; see also the quote of Nobel prize laureate Wigner [121] at the beginning of this section.

The problem with releasing power is that it can be used for good and for bad purposes. Science is objective in the sense that a certain theory is either consistent with observations, or it is not. However, scientific knowledge does not come with the moral standard that tells us *how* the power that we release in our scientific efforts should be used. It is essential that each of us develops such a standard, so that the fruits of our knowledge can be used for the benefit of mankind and the world we inhabit.

References

[1] Abramowitz, M. and I. A. Stegun, 1965, *Handbook of Mathematical Functions*, New York, Dover Publications.

[2] Aharonov, Y., and D. Bohm, 1959, Significance of electromagentic potentials in the quantum theory, *Phys. Rev.*, **115**, 485–491.

[3] Aki, K. and P. G. Richards, 2002, *Quantitative Seismology*, 2nd edition, Sausalito, University Science Books.

[4] Aldridge, D. F., 1994, Linearization of the eikonal equation, *Geophysics*, **59**, 1631–1632.

[5] Arfken, G. B., 1995, *Mathematical Methods for Physicists*, San Diego, Academic Press.

[6] Ashby, N., 2002, Relativity and the Global Positioning System, *Phys. Today*, **55(5)**, 41–47.

[7] Backus, M. M., 1959, Water reverberations – their nature and elimination, *Geophysics*, **24**, 233–261.

[8] Barish, B. C. and R. Weiss, 1999, LIGO and the detection of gravitational waves, *Phys. Today*, **52(10)**, 44–55.

[9] Barton, G., 1989, *Elements of Green's Functions and Propagation, Potentials, Diffusion and Waves*, Oxford, Oxford Scientific Publications.

[10] Beissner, K., 1998, The acoustic radiation force in lossless fluids in Eulerian and Lagrangian coordinates, *J. Acoust. Soc. Am.*, **103**, 2321–2332.

[11] Bennett, J. A., 1973, Variations of the ray path and phase path: a Hamiltonian formulation, *Radio Sci.*, **8**, 737–744.

[12] Bevington P. R. and D. K. Robinson, 1992, *Data reduction and Error Analysis for the Physical Sciences*, New York, McGraw-Hill.

[13] Bellman, R. and R. Kalaba, 1960, Invariant embedding and mathematical physics I. Particle processes, *J. Math. Phys.*, **1**, 280–308.

[14] Bender, C. M. and S. A. Orszag, 1978, *Advanced mathematical Methods for Scientists and Engineers*, New York, McGraw-Hill.

[15] Berry, M. V. and S. Klein, 1997, Transparent mirrors: rays, waves and localization, *Eur. J. Phys.*, **18**, 222–228.

[16] Berry, M. V. and C. Upstill, 1980, Catastrophe optics: Morphologies of caustics and their diffraction patterns, *Prog. Optics*, **18**, 257–346.

[17] Blakeley, R. J., 1995, *Potential Theory in Gravity and Magnetics*, Cambridge, Cambridge University Press.

[18] Bleistein, N. and R. A. Handelsman, 1975, *Asymptotic Expansions of Integrals*, New York, Dover.

[19] Boas, M. L., 1983, *Mathematical Methods in the Physical Sciences*, 2nd edition, New York, Wiley.

[20] Brack, M. and R. K. Bhaduri, 1997, *Semiclassical Physics*, Reading MA, Addison-Wesley.

[21] Brillouin, L., 1960, *Wave Propagation and Group Velocity*, New York, Academic Press.

[22] Broglie, L. de, 1952, *La Théorie des particules de spin 1/2*, Paris, Gauthier-Villars.

[23] Buckingham, E., 1914, On physically similar systems; illustrations of the use of dimensional equations, *Phys. Rev.*, **4**, 345–376.

[24] Butkov, E., 1968, *Mathematical Physics*, Reading MA, Addison Wesley.

[25] Choy, G. L., and P. G. Richards, 1975, Pulse distortion and Hilbert transformation in multiply reflected and refracted body waves, *Bull. Seismol. Soc. Am.*, **65**, 55–70.

[26] Cipra, B., 2000, *Misteaks . . . and how to Find Them Before the Teacher Does*, 3rd Edition, Natck MA, A. K. Peters.

[27] Claerbout, J. F., 1976, *Fundamentals of Geophysical Data Processing*, New York, McGraw-Hill.

[28] Claerbout, J. F., 1985, *Imaging the Earth's Interior*, Oxford, Blackwell.

[29] Chopelas, A., 1996, Thermal expansivity of lower mantle phases MgO and $MgSiO_3$ perovskite at high pressure derived from vibrational spectroscopy, *Phys. Earth. Plan. Int.*, **98**, 3–15.

[30] Coveney, P., and R. Highfield, 1991, *The Arrow of Time*, London, Harper Collins.

[31] Dahlen, F. A., 1979, The spectra of unresolved split normal mode multiples, *Geophys. J. R. Astron. Soc.*, **58**, 1–33.

[32] Dahlen, F. A. and I. H. Henson, 1985, Asymptotic normal modes of a laterally heterogeneous Earth, *J. Geophys. Res.*, **90**, 12653–12681.

[33] Dahlen, F. A., and J. Tromp, 1988, *Theoretical Global Seismology*, Princeton, Princeton University Press.

[34] DeSanto, J. A., 1992. *Scalar Wave Theory; Green's Functions and Applications*, Berlin, Springer Verlag.

[35] Dziewonski, A. M., and J. H. Woodhouse, 1983, Studies of the seismic source using normal-mode theory, in *Earthquakes: Observations, Theory and Interpretation*, edited by H. Kanamori and E. Boschi, Amsterdam, North Holland, pp. 45–137.

[36] Earnshaw, S., 1842, On the nature of molecular forces which regulate the constitution of the luminiferous ether, *Trans. Camb. Phil. Soc.*, **7**, 97–112.

[37] Edmonds, A. R., 1974, *Angular Momentum in Quantum Mechanics*, 3rd edition, Princeton, Princeton University Press.

[38] Farra, V. and R. Madariaga, 1987, Seismic waveform modeling in heterogeneous media by ray perturbation theory, *J. Geophys. Res.*, **92**, 2697–2712.

[39] Feynman, R. P., 1975, *The Character of Physical Law*, Cambridge (MA), MIT Press.

[40] Feynman, R. P. and A. R. Hibbs, 1965, *Quantum Mechanics and Path Integrals*, New York, McGraw-Hill.

[41] Fishbach, E. and C. Talmadge, 1992, Six years of the fifth force, *Nature*, **356**, 207–214.

[42] Fletcher, C., 1996, *The Complete Walker III*, New York, Alfred A. Knopf.

[43] Fokkema, J. T. and P. M. van den Berg, 1993, *Seismic Applications of Acoustic Reciprocity*, Amsterdam, Elsevier.

[44] Goldstein, H., 1980, *Classical Mechanics*, 2nd edition, Reading MA, Addison-Wesley.

[45] Gradshteyn, I. S. and I. M. Rhyzik, 1965, *Tables of Integrals, Series and Products*, New York, Academic Press.

[46] Gubbins, D. and R. Snieder, 1991, Dispersion of P waves in subducted lithosphere: Evidence for an eclogite layer, *J. Geophys. Res.*, **96**, 6321–6333.

[47] Guglielmi, A. V. and O. A. Pokhotelov, 1996, *Geoelectromagnetic Waves*, Bristol, Inst. of Physics Publ.

[48] Halbwachs, F., 1960, *Théorie relativiste des fluides a spin*, Paris, Gauthier-Villars.

[49] Hildebrand, A. R., M. Pilkington, M. Conners, C. Ortiz-Aleman and R. E. Chavez, 1995, Size and structure of the Chicxulub crater revealed by horizontal gravity and cenotes, *Nature*, **376**, 415–417.

[50] Holton, J. R., 1992, *An Introduction to Dynamic Meteorology*, San Diego, Academic Press.

[51] Iyer, H. M. and K. Hirahara (eds.), 1993, *Seismic Tomography, Theory and Practice*, London, Prentice Hall.

[52] Ishimaru, A., 1997, *Wave propagation and Scattering in Random Media*, Oxford, Oxford University Press.

[53] Jackson, J. D., 1975, *Classical Electrodynamics*, New York, Wiley.

[54] Jeffreys, H., 1924, On certain approximate solutions of linear differential equations of second order, *Phil. London Math. Soc.*, **23**, 428–436.

[55] Kermode, A. C., 1996, *Mechanics of Flight*, 10th edition, Singapore, Longman.

[56] Kline, S. J., 1965, *Similitude and Approximation Theory*, New York, McGraw-Hill.

[57] Kravtsov, Yu. A., 1988, Ray and caustics as physical objects, *Prog. Optics*, **26**, 228–348.

[58] Kulkarny, V. A., and B. S. White, 1982, Focusing of waves in turbulent inhomogeneous media, *Phys. Fluids*, **25**, 1770–1784.

[59] Lambeck, K., 1988, *Geophysical Geodesy*, Oxford, Oxford University Press.

[60] Lanczos, C., 1970, *The Variational Principles of Mechanics*, New York, Dover.

[61] Lauterborn, W., T. Kurz and M. Wiesenfeldt, 1993, *Coherent Optics*, Berlin, Springer Verlag.

[62] Lavenda, B. H., 1991, *Statistical Physics, a Probabilistic Approach*, New York, John Wiley.

[63] Lin, C. C., L. A. Segel and G. H. Handelman, 1974, *Mathematics Applied to Deterministic Problems in the Natural Sciences*, New York, Macmillan.

[64] Lister, G. S., and P. F. Williams, 1983, The partitioning of deformation in flowing rock masses, *Tectonophysics*, **92**, 1–33.

[65] Love, S. G., and D. E. Brownlee, 1993, A direct measurement of the terrestrial mass accretion rate of cosmic dust, *Science*, **262**, 550–553.

[66] Madelung, E., 1926, Quantentheorie in hydrodynamischer form, *Z. Phys.*, **40**, 322.

[67] Marchaj, C. A., 1993, *Aerohydrodynamics of Sailing*, 2nd edition, London, Adlard Coles Nautical.

[68] Marsden, J. E. and A. J. Tromba, 1988, *Vector Calculus*, New York, Freeman and Company.

[69] Merzbacher, E., 1970, *Quantum Mechanics*, New York, Wiley.

[70] Moler, C. and C. van Loan, 1978, Nineteen dubious ways to compute the exponential of a matrix, *SIAM Review*, **20**, 801–836.

[71] Morley, T., 1985, A simple proof that the world is three dimensional, *SIAM Review*, **27**, 69–71.

[72] Morrison, L. and R. Stephenson, 1988, The sands of time and the Earth's rotation, *Astronomy and Geophysics*, **39(5)**, 8–13.

[73] Muirhead, H., 1973, *The Special Theory of Relativity*, London, Macmillan.

[74] Nayfeh, A. H., 1981, *Introduction to Perturbation Techniques*, New York, Wiley.

[75] Nolet G. (ed.), 1987, *Seismic Tomography, with Applications in Global Seismology and Exploration Geophysics*, Dordrecht, Kluwer.

[76] Ohanian, H. C. and R. Ruffini, 1994, *Gravitation and Spacetime*, New York, Norton.

[77] Olson, P., 1989, Mantle convection and plumes, in *The Encyclopedia of Solid Earth Geophysics*, ed. D. E. James, New York, Van Nostrand Reinholt.

[78] Oort, A. H. and J. P. Peixoto, 1992, *Physics of Climate*, New York, Springer Verlag.

[79] Parker, R. L., 1994, *Geophysical Inverse Theory*, Princeton NJ, Princeton University Press.

[80] Parker, R. L. and M. A. Zumberge, 1989, An analysis of geophysical experiments to test Newton's law of gravity, *Nature*, **342**, 29–31.

[81] Parsons, B. and J. G. Sclater, 1977, An analysis of the variation of the ocean floor bathymetry and heat flow with age, *J. Geophys. Res.*, **32**, 803–827.

[82] Pedlosky, J., 1979, *Geophysical Fluid Dynamics*, Berlın, Springer Verlag.

[83] Popper, K., 1956, The arrow of time, *Nature*, **177**, 538.

[84] Press, W. H., B. P. Flannery, S. A. Teukolsky and W. T. Vetterling, 1986, *Numerical Recipes*, Cambridge, Cambridge University Press.

[85] Price, H., 1996, *Time's Arrow and Archimedes' Point, New Directions for the Physics of Time*, New York, Oxford University Press.

[86] Rayleigh, Lord, 1917, On the reflection of light from a regularly stratified medium, *Proc. Roy. Soc. Lon.*, **A93**, 565–577.

[87] Riley, K. F., M. P. Hobson, and S. J. Bence, 1998, *Mathematical Methods for Physics and Engineering*, Cambridge, Cambridge University Press.

[88] Rossing, T. D., 1990, *The Science of Sound*, Reading (MA), Addison Wesley.

[89] Rummel, R., 1986, Satellite gradiometry, in *Mathematical Techniques for High-resolution Mapping of the Gravitational Field*, Lecture notes in the Earth Sciences, vol. 7, ed. H. Suenkel, Berlin, Springer Verlag.

[90] Robinson, E. A. and S. Treitel, 1980, *Geophysical Signal Analysis*, Englewood Cliffs (NJ), Prentice Hall.

[91] Sakurai, J. J., 1978, *Advanced Quantum Mechanics*, Reading (MA), Addison Wesley.

[92] Scales, J. A. and R. Snieder, 1997, Humility and nonlinearity, *Geophysics*, **62**, 1355–1358.

[93] Schneider, W. A., 1978, Integral formulation for migration in two and three dimensions, *Geophysics*, **43**, 49–76.

[94] Silverman, M. P., 1993, *And Yet It Moves; Strange Systems and Subtle Questions in Physics*. Cambridge, Cambridge University Press.

[95] Smith, D. R., 1974, *Variational Methods in Optimization*, Englewood Cliffs, Prentice-Hall.

[96] Snieder, R. K., 1985, The origin of the 100,000 year cycle in a simple ice age model, *J. Geophys. Res.*, **90**, 5661–5664.

[97] Snieder, R., 1986, 3D Linearized scattering of surface waves and a formalism for surface wave holography, *Geophys. J. R. astr. Soc.*, **84**, 581–605.

[98] Snieder, R., 1996, Surface wave inversions on a regional scale, in *Seismic Modelling of Earth Structure,* eds. E. Boschi, G. Ekstrom and A. Morelli, Bologna, Editrice Compositori, pp. 149–181.

[99] Snieder, R. and D. F. Aldridge, 1995, Perturbation theory for travel times, *J. Acoust. Soc. Am.*, **98**, 1565–1569.

[100] Snieder, R., 2002, Time-reversal invariance and the relation between wave chaos and classical chaos, in *Imaging of Complex Media with Acoustic and Seismic Waves*, pp. 1–15, eds. M. Fink, W. A. Kuperman, J. P. Montagner, and A. Tourin, Berlin, Springer Verlag.

[101] Snieder, R. and G. Nolet, 1987, Linearized scattering of surface waves on a spherical Earth, *J. Geophys.*, **61**, 55–63.

[102] Spetzler, J. and R. Snieder, 2001, The formation of caustics in two and three dimensional media, *Geophys. J. Int.*, **144**, 175–182.

[103] Stacey, F. D., 1992, *Physics of the Earth*, 3rd edition, Brisbane, Brookfield Press.

[104] Tabor, M., 1989, *Chaos and Integrability in Nonlinear Dynamics*, New York, John Wiley.

[105] Taylor, E. F., and J. A. Wheeler, 1966, *Spacetime Physics*, San Francisco, Freeman.

[106] Tennekes, H., 1997, *The Simple Science of Flight: from Insects to Jumbo Jets*, Cambridge MA, MIT Press.

[107] Thompson, P. A., 1972, *Compressible-fluid Dynamics*, New York, McGraw-Hill.

[108] Tritton, D. J., 1982, *Physical Fluid Dynamics*, Wokingham (UK), Van Nostrand Reinhold.

[109] Tromp, J. and R. Snieder, 1989, The reflection and transmission of plane P- and S-waves by a continuously stratified band: a new approach using invariant embedding, *Geophys. J.*, **96**, 447–456.

[110] Turcotte, D. L. and G. Schubert, 2002, *Geodynamics*, 2nd edition, Cambridge UK, Cambridge University Press.

[111] van Dyke, M., 1978, *Perturbation Methods in Fluid Mechanics*, Stanford CA, Parabolic Press.

[112] Virieux, J., 1996, Seismic ray tracing, in *Seismic Modelling of Earth Structure*, eds. E. Boschi, G. Ekstrom and A. Morelli, Bologna, Editrice Compositori, pp. 221–304.

[113] Vogel, S., Exposing life's limits with dimensionless numbers, *Physics Today*, **51(11)**, 22–27.

[114] Udias, A., 1999, *Principles of Seismology*, Cambridge, Cambridge University Press.

[115] Watson, T. H., 1982, A real frequency, wave-number analysis of leaking modes, *Bull. Seismol. Soc. Am.*, **62**, 369–394.

[116] Webster, G. M. (ed.), 1981, *Deconvolution*, Geophysics reprint series, vol. 1, Tulsa, Society of Exploration Geophysicists.

[117] Weisskopf, V. F., 1939, On the self-energy and the electric field of the electron, *Phys. Rev.*, **56**, 72–85.

[118] Whitaker, S., 1968, *Introduction to Fluid Mechanics*, Englewood Cliffs, Prentice-Hall.

[119] Whitham, G. B., 1974, *Linear and Nonlinear Waves*, New York, Wiley.

[120] Wigner, E. P., 1960, The unreasonable effectiveness of mathematics in the natural sciences, *Comm. Pure Appl. Math.*, **13**, 222–236.

[121] Wigner, E. P, 1972, The place of consciousness in modern physics, in *Consciousness and reality*, eds. C. Muses and A. M. Young, New York, Outerbridge and Lizard, pp. 132–141.

[122] Wu, R. S. and K. Aki, 1985, Scattering characteristics of elastic waves by an elastic heterogeneity, *Geophysics*, **50**, 582–595.

[123] Yilmaz, O., 1987, *Seismic Data Processing, Investigations in Geophysics*, vol. 2, Tulsa, Society of Exploration Geophysicists.

[124] Yoder, C. F., J. G. Williams, J. O. Dickey, B. E. Schutz, R. J. Eanes and B. D. Tapley, 1983, Secular variation of the Earth's gravitational harmonic J_2 coefficient from LAGEOS and nontidal acceleration of the Earth rotation, *Nature*, **303**, 757–762.

[125] Zee, A., 2003, *Quantum Field Theory in a Nutshell*, Princeton NJ, Princeton University Press.

[126] Ziolkowski, A., 1991, Why don't we measure seismic signatures? *Geophysics*, **56**, 190–201.

[127] Zumberge, M. A., J. R. Ridgway and J. A. Hildebrand, 1997, A towed marine gravity meter for near-bottom surveys, *Geophysics*, **62**, 1386–1393.

Index

acceleration in spherical coordinates 37–40
accretion 22
acoustic impedance 432
action 476
advanced Green's function 304
advection 140
advective terms 159–162
Aharonov–Bohm effect 104–108
Airy phase 453
analytic functions 21
aquifers 13
analytic continuation 250
analytic function 247
  and fluid flow 251
angle of attack 11
angular momentum 5
anti-symmetric tensor 402
Archimedes's law 168
arrow of time 266, 304
associated Legendre functions 319–321
  asymptotic behavior 331
  orthogonality 326
asymptotic behavior
  associated Legendre functions 331
  Bessel function 301, 327–329, 455–456
  Hankel function 301
  Legendre functions 330–334
  Neumann function 301
asymptotic evaluation of integrals
  integration by parts 438–440
  Taylor series 437–438
asymptotic expansion 417, 444
augmented matrix 198
averaging integrals 130–132
Avogadro's number 445

backsubstitution 198
basis vectors
  Cartesian 34
  spherical 34
Bessel equation 300, 315, 327
Bessel function 300, 315
  asymptotic behavior 301, 327–329, 455–456
  integral representation 360, 453–455
  orthogonality 325
  zero crossings 316–317
β-plane approximation 184
blue sky 165, 421
Bohr–Sommerfeld quantization rule 332
Born approximation 29, 417–421
bouncing ball 24–27
boundary condition
  homogeneous 284
  periodic 315
  radiation 299
boundary layer problems 436
boxcar function 203
Brunt-Väisälä frequency 10
Buckingham pi theorem 7–10
buoyancy force 168

Cartesian basis vectors 34
Cartesian coordinates 31, 32
  relation to cylindrical coordinates 44
  relation to spherical coordinates 33
catenary 489
Cauchy–Riemann relations 247
causal
  filter 120
  Green's function 274
  response 338
causality 265
  and dissipation 266
causal solution 268, 303
caustic 166, 436
centrifugal force 75, 154, 179–184
chain line 487–491
changing coordinate systems 35
chaos theory 153
chaotic system 153
charge density 70
Chicxulub impact crater 354

circulation 109
classical mechanics 8, 17
  and Planck's constant 294
  and ray theory 481
  variational formulation 476–477
closure relation 176, 225
colatitude 31
complete set 176, 186, 218, 335
completeness relation 186
concave function 116
contraction
  definition 398
  double 399
conservation law
  energy of continuous medium 138–139
  general form 133–135
  heat 141
  mass of continuous medium 135–136, 150
  momentum of continuous medium 138–139, 151
consistency 164
constrained optimization 481
  chain line 487–491
  design of a can 485–487
constructive interference 106
continuity equation 135–136, 150
convection 140
  Earth's mantle 167–169
convergence 20
convex function 116
convolution 229
  Fourier transform 228–230
  linear filter 233
  three dimensional 273
convolution theorem 229, 231–234
coordinate transform
  matrix 391–393
  time-dependent 395
  vector 379–382
Coriolis force 154, 179–184
Coriolis parameter 184
correlation 230
  Fourier transform 230–231
coupling constant 70, 85, 184
curl
  cylindrical coordinates 87
  definition 78
  relation to line integral 79
  rigid rotation 80–83
  shear 83–84
  spherical coordinates 86–87
current
  electric 85, 102
  general 133
  heat 140
  kinetic energy 139
  momentum 137
  potential energy 139
  probability 95
  small-scale momentum 148
curvature of a function 113–117
curvilinear coordinate systems 34, 37
cylindrical coordinates 43–45
  curl 87
  divergence 71–73
  gradient 63
  Laplacian 130
  relation to Cartesian coordinates 44

d'Alembert solution 349
de Broglie relation 96
deconvolution 230, 236
deformation 386
degenerate modes 320
delta function 202–206
  Fourier transform 224–225, 289
  Green's function 271
  input for linear filter 232
  more dimensions 210, 289
  of a function 208–210
  on a sphere 210–212
  point charge or mass 212–216
  properties 206–208
density
  charge 70
  mass 70
  probability 94
dereverberation 236
  filter 234–238
derivative
  directional 54
  estimation of 156–159
  partial 57–61, 135
  second, and radius of curvature 115
  total 57–61, 135
design criteria 465
destructive interference 106
diffusion constant 294
dimensionless
  equation 169–172
  numbers 4, 169
dipole
  field 371
  gravitational or electric 365–369
  moment 367
  vector 367
Dirac delta function, *see* delta function
directional derivative 54
direct product 400
dispersion 343, 453
  relation 343
displacement vector 401
distribution 205
divergence
  cylindrical coordinates 71–73
  definition 67
  electrical field 70
  gravitational field 73
  magnetic field 71
  relation to flux 67
  spherical coordinates 73
  velocity field 136

drum
  normal modes 314–317
dyad 175, 400
dynamo 103

$E = mc^2$ 410
Earnshaw's theorem 127
Earth
  moment of inertia 375
  normal mode spectrum 337
  quadrupole field 374–377
  temperature Green's function 279–284
earthquake 160
eigenfrequency 191
eigenfunction 335
  translation operator 243
eigenvalue 185, 242, 335
eigenvalue decomposition of matrix 184–187
eigenvector 185, 241–242
  decomposition 192, 240
eikonal equation 165, 422, 480
Einstein summation convention 389–391
elasticity tensor 406
electrical field
  divergence 67
  energy 126
  flux 65
  Green's function 272–273
  relation to gravity field 354
electromotive force 104
energy
  conservation and Newton's law 55–57
  self 212–216
  static electric field 126
equation
  Bessel 300, 315, 327
  continuity 135–136, 150
  dimensionless 169–172
  eikonal 165, 422, 480
  Euler–Lagrange 472–474
  heat 141, 167, 280
  Helmholtz 297, 312, 340, 417
  homogeneous 278, 282, 341
  inhomogeneous 278
  kinematic ray tracing 166, 480
  Laplace 122, 127, 129, 262, 447
  Maxwell 103
  Navier–Stokes 15, 149, 168
  Poisson 125–126, 354
  Schrödinger 5
  transcendental 343, 489
  transport 166
  Verhulst 284
  wave 302, 349
equilibrium
  point 116
  stable and unstable 116
escape velocity 411
estimation of derivative 156–159
ethics 493

Eulerian formulation 61, 135
  equation of motion 136
Euler–Lagrange equation 472–474
evanescent waves 342
expansion coefficient 168
explosion of nuclear bomb 145
extremum 114

factorial $n!$ 441
  behavior for large n 441–445
Fermat's theorem 424
Fick's law 141
field lines
  magnetic 66
fifteen degrees approximation 115
fifth force 377–378
filter
  dereverberation 234–238
  linear 231–232
finding mistakes 6–7
filters
  causal 240
  frequency 238–240
  linear 231–232, 264
  low-pass 238
flux 64–66
  definition 64
  electric field 65
  magnetic field 65, 108
  relation to divergence 67
force
  buoyancy 168
  centrifugal 75, 154, 179–184
  Coriolis 154, 179–184
  fifth 377–378
  gravitational 75
  pressure 50–53, 154
  quantum 152
Fourier transform
  complex, over finite interval 221–223
  convolution 228–230
  correlation 230–231
  delta function 224–225, 289
  infinite interval 223–224
  real, over finite interval 217–221
  shift property 236
  sign and scale factor 225–228
four-vector 408
f-plane approximation 184
frequency filter 238–240
function
  analytic 21, 247
  associated Legendre 319–321
  Bessel 300, 315, 360
  boxcar 203
  concave and convex 116
  curvature of 113–117
  Gaussian 206, 292, 439–440, 443
  gamma 441
  Hankel 300
  harmonic 130

  Legendre 320
  Neumann 300, 315
  nonanalytic 254–255
  of matrix 187–189
  sinc 239
functional 464
fundamental mode 313

gamma function 441
  recursive relation 441
  relation to $n!$ 441
Gateway Arch of St. Louis 487–491
Gauss function 206, 292, 439–440, 443
Gauss's law 88–89
  relation to Stokes's law 100–101
geometric ray theory, *see* ray theory
global bahavior of a function 20
Global Positioning System 411
  and relativistic effects 411
gradient
  cylindrical coordinates 63
  definition 48
  direction 49
  magnitude 50
  of vector 401
  properties 46–50
  spherical coordinates 61–63
gravity anomaly 353
gravitational field
  divergence 73
  Green's function 354–356
  in spherical harmonics 361–364
  measured from aircraft or satellite 356
  N-dimensions 74–75
  relation to electrical field 354
  spherically symmetric mass 89–91
gravitational waves 369
great-circle 468–472
Green's function
  advanced 304
  causal 274, 303
  Earth's temperature 279–284
  electrical field 272–273
  general problem 276–279
  gravitational potential 354–356
    in spherical harmonics 361–364
  harmonic oscillator 267–271
  heat equation in N dimensions 288–292
  Helmholtz equation in 1D 296–298
  Helmholtz equation in 2D and 3D 298–302
  impulse response 271, 273–276
  nonlinear system 284–287
  normal modes 334–340
  reciprocity 308–309
  response to delta function excitation 213, 271
  retarded 303
  Schrödinger equation 292–296
  singularity 276
  wave equation in 1D 307–308
  wave equation in 2D 304–307
  wave equation in 3D 302–303
group velocity 344, 451
  stationary phase 450–453
guided waves 340–344

Hankel function 300
  asymptotic behavior 301
harmonious spectra 320
harmonic function 130, 247, 249
harmonic oscillator 7
  Green's function 267–271
  perturbation analysis 425
heat current 140
heat equation 141, 167, 280
  Green's function in N dimensions 288–292
Heisenberg uncertainty relation 239
Helmholtz equation 297, 312, 340, 417
  Green's function in 1D 296–298
  Green's function in 2D and 3D 298–302
Hermitian
  conjugate 409
  operator 334
Hessian 399
  relation to Laplacian 400
hidden numerical factors 159
higher modes 313
homogeneous
  boundary conditions 284
  equation 278, 282, 341
Householder transformation 197–201
Huygens's principle 93
hydraulic jump 161
hydrocarbon reservoirs 13
hydrodynamics and quantum mechanics 150–152

idempotent operator 174
identity
  matrix 202
  operator 176
ill-conditioned 197
imaging 93
impedance 432
  acoustic 432
impulse response
  definition 232
  Green's function 271, 273–276
incompressible fluid 67
induced drag 111
inductance 104
inertia tensor 372
inhomogeneous equation 278
instability
  numerical 197
  of matter 126–128
integrable singularity 302
integration by parts 82
  asymptotic evaluation of integrals 438–440
integration without primitive function 259–262
invariant embedding 30

inverse
  matrix 187
  problems 192, 197

Jacobian 41, 212

kinetic energy 25, 55, 476
  current 139
kinematic ray tracing 480
Kronecker delta 137, 185, 202, 383

Lagrange multipliers
  discrete problem 481–484
Lagrangian 476
Lagrangian formulation 61, 135
  classical mechanics 476–477
  equation of motion 136
Laplace equation 122, 127, 129, 262, 447
Laplacian
  cylindrical coordinates 130
  definition 122
  invariance 400
  relation to Hessian 400
  spherical coordinates 130
  unit sphere 362
latitude 31
layered media 29
leaky modes 344–348
Legendre functions 320
  asymptotic behavior 330–334
Lenz's law 103–104
lift coefficient 11
lift of a wing 11–12, 108–110
lightning 128–129
linear filter 231–232, 358
linear spring 5
linear system 270
line integral
  relation to curl 79
local wavenumber 295, 430
longitude 31
Lorentz transformation 408
Love waves 343
low-pass filter 238

magnetic field
  divergence 71
  flux 66, 104
  lines 65
  straight current 85, 102
magnetic induction 103–104
magnetic monopoles 71, 368
mantle convection 167–169
mass density 70
mathematical biology 285
matrix
  augmented 198
  coordinate transform 391–393
  eigenvalue decomposition 184
  function of 187–189
  identity 202
  inverse 187
  inverse of orthogonal 36
  inverse of unitary 385
  orthogonal 36
  rotation 385
  trace of 390, 399
  transpose of 36, 175, 384
  unitary 185, 382–385
Maxwell equation 103
meteorites 22
method of stationary phase 450–453
  group velocity 450
method of steepest descent 445–450
  multidimensional integrals 449
minimax arrivals 480
moment of inertia
  Earth 375
momentum current 137
momentum-energy vector 410
monopole
  field 371
  magnetic 71, 368
  gravitational or electric 365–369
multiple scale analysis 427
multipole expansion 369–374

n! 441
  behavior for large *n* 441–445
Navier–Stokes equation 15, 149, 168
Neumann function 300, 315
  asymptotic behavior 301
Neumann series 420
New Trier Connections Project 490
nonanalytic functions 254–255
nonlinear system
  Green's function 284–287
nonuniqueness 90
normal modes
  degenerate 320
  drum 317–318
  Earth 337
  fundamental 313
  Green's function 334–340
  guided waves 340–344
  higher modes 313
  leaky 344–348
  orthogonality relation 323–327
  plate 312
  sphere 317–322
  string 312–314
  summation 339
  vibrating system 189–192
nuclear chain reaction 145
numerical solution
  differential equation 158
  errors 23
  instability 198
  stability 158

Ohm's law 432
operator
  Hermitian 334

  identity 176
  translation 243
optimization problem 463
  constrained 482
  design of a can 461–463
  shape of a can 463–465
orthogonality
  associated Legendre functions 326
  Bessel functions 325
  modes of string 313
  normal modes 323–327
  spherical harmonics 327
orthogonal matrix 36
  inverse of 36
orthonormal
  basis 34
  coordinate system 382
  transformation 101
overtones 313, 320

Parsevals's theorem 231
partial derivative 57–61, 135
particular solution 278
periodic boundary condition 315
perturbation
  singular 172
perturbation theory
  limits on 424–427
  need for consistency 431–433
  regular 413–417
  singular 433–436
phase velocity 451
Physical Acoustic Laboratory 312
physical dimension 3
pipe flow 13–15
Planck's constant 94, 151
  and classical mechanics 294
  and diffusion constant 294
Planck's law 445
point charge or mass 212
point of stationary phase 451
Poisson's equation 125, 354
polar and apolar substances 367
pole 254
  simple 258
poorly conditioned 188
population growth 285
potential energy 5, 55, 476
  current 139
potential flow 251
potential temperature 10
power 57, 139
pressure
  force 50–53, 154
  gradient 15
  isotropic property 406–407
probability 439
  Gauss distribution 439–440
probability density 94
probability-density current 95
projection
  Fourier analysis 219
  linear algebra 173–177
  on plane 178
pseudo-tensor 398

quadrupole
  field 372–373
  field of Earth 374–377
  gravitational or electric 365–369
quantum field theory 215
quantum force 152
quantum mechanics 94, 106, 127, 332
  relation to hydrodynamics 150–152
  WKB approximation 431

radiation boundary condition 299
radiation damping 348–352
radiation stress 138
radiogenic heating 280
radius of convergence 20
radius of curvature 114
  and second derivative 115
rainbow 166
ray 166
  as curve of stationary travel time 478–481
Rayleigh
  number 169, 171
  waves 343, 453
ray-geometric reflection point 458
ray theory 162–167, 458, 478
  and classical mechanics 481
  equation of kinematic ray tracing 166
  frequency dependence 165
  transport equation 166
reciprocity theorem 309
reflected waves
  stationary phase method 456–460
reflection and transmission 27–30
reflection coefficient 27
regular perturbation theory 413–417
representation theorem
  acoustic waves 91–93
residue 264
  definition 257
  simple pole 258
residue theorem 255–258
resonance 192
retarded Green's function 303
reverberation 29
Rossby waves 184
rotation matrix 385

saddle point 448
saddle surface 124
scalar 393
scale factors 34, 41
scaling relations 12–13
schlieren method 161
science of flight 13
Schrödinger equation 5, 94
  Green's function 292–296
secular growth 426
self energy 212–216

separation of variables 314, 318
shear traction 404, 406
shift property of Fourier transform 236
shock waves 161
shortest distance
  in a plane 117–120, 465–468
  on a sphere 468–472
simple pole
  definition 258
  residue 258
sinc function 239
single scattering 421
singular
  Green's function 276
  perturbation 172
  perturbation theory 433–436
  value decomposition 192–197
small-scale momentum current 148
Snell's law 458
soap film 120–124
sources and sinks 69–71, 133
sources of curvature 124–126
special relativity 408–411
spectrum
  Earth 337
  harmonious 320
  saxophone 218
sphere
  normal modes 317–322
spherical basis vectors 34
spherical coordinates 31–34
  acceleration 37–40
  definition 31
  curl 86–87
  divergence 73
  gradient 61–63
  Laplacian 129
  relation to Cartesian coordinates 33
  velocity 38
  volume integration 40–45
spherical harmonics 320, 362
  orthogonality 327
spreading of pollution 61
stability
  equilibrium point 116
  of gravitational motion 127
  of matter 126–128
  of numerical solution 158
  of planetary orbits 73–77
stationary point 114, 446, 451, 458
stationary phase method 450–453
  Bessel function 455–456
  group velocity 450
  reflected waves 456–460
statistical mechanics 441, 445
steepest descent method 445
  multidimensional integrals 449
Stirling's formula 444
strain tensor 401–403
stress tensor 403–406
  in fluid 406
Stokes's law 97–100, 248
  relation to theorem of Gauss 100–101
streamlines 69
string
  normal modes 312–314
summation convention 389–391
surface waves 334
superposition principle 270, 284, 286
surface element 42
surface tension 120
SVD 196
symmetric tensor 402

tangent circle 114
Taylor series 16–22
  asymptotic valuation of integrals 437–438
  function of more variables 19
  of common functions 18
tensor
  antisymmetric 402
  definition 393–396
  elasticity 406
  inertia 372
  non-Cartesian 398
  pseudo 398
  rank 0, 1, 2, *n* 394
  rank 4 406
  strain 401–403
  stress 403–406
  symmetric 402
theorem
  Buckingham pi 7–10
  Cauchy–Riemann 245–249
  convolution 229
  Earnshaw 127
  Fermat's 424
  Gauss 88–89
  Parseval 231
  reciprocity 309
  representation 91–93
  residue 255–258
  Stokes 97–100, 248
thermal expansion coefficient 168
tomography 421–424
total derivative 57–61, 135
trace of matrix 390
  invariance 399
traction 404
transcendental equation 343, 489
translation operator 243
  eigenfunctions 243
transmission coefficient 27
transparent mirrors 30
transport equation 166
transpose
  of a vector 175
  of a matrix 36, 175, 384
travel time 165
  tomography 421–424
truncation error 416
two-slit experiment 106–107

uncertainty relation of Heisenberg 239
unit, physical 3
unitary
  matrix 185, 382–385
  transformation 383
unstable equilibium 116
upper-triangular form 198
upward continuation
  flat geometry in 2D 356–359
  flat geometry in 3D 359–361

variational calculus 117, 464
vector
  coordinate transform 379–382
  four- 408
  momentum-energy 410
  momentum-potential 105
  transpose of 175
velocity
  estimation 93
  group 344, 451
  phase 451
velocity in spherical coordinates 37
Verhulst equation 284
vibrating system
  normal modes 189–192
viscosity 14, 149
volume integration in spherical coordinates 40–45
volume of a sphere 40
vortex in bathtub 154–156
vortex strength 110
vorticity
  definition 80
  rigid rotation 81–83
  shear 83–84

wave equation 302, 349
  Green's function in 1D 307–308
  Green's function in 2D 304–307
  Green's function in 3D 302–303
wavenumber
  local 295, 430
waves
  dispersive 343
  electromagnetic 343
  evanescent 342
  gravitational 369
  guided 340–344
  Love 343
  Rayleigh 343, 453
  Rossby 184
  shock 161
  surface 334
weather map 50
weather prediction 147
wingtip vortices 108–112
WKB approximation 427–431
  acoustic waves 430
  quantum mechanics 431

zero measure 203

TORONTO PUBLIC LIBRARY

A Guided Tour of
Mathematical Methods for the Physical Sciences

Mathematical methods are essential tools for all physical scientists. This second edition of *A Guided Tour of Mathematical Methods* provides a comprehensive tour of the mathematical knowledge and techniques that are needed by students in this area. In contrast to more traditional textbooks, all the material is presented in the form of problems. Within these problems the basic mathematical theory and its physical applications are well integrated. The mathematical insights that the student acquires are therefore driven by their physical insight. Topics that are covered include vector calculus, linear algebra, Fourier analysis, scale analysis, complex integration, Green's functions, normal modes, tensor calculus, and perturbation theory. The second edition contains new chapters on dimensional analysis, variational calculus, and the asymptotic evaluation of integrals. This book can be used by undergraduates and lower-level graduate students in the physical sciences. It can serve as a stand-alone text, or as a source of problems and examples to complement other textbooks.

Quotes from reviews of the first edition

> "... a splendid book, quite a delight to see mathematical models from a different perspective ... treatment is so smooth that it is hard to know at any instant whether one is learning new tools or assimilating the applications to interesting examples ... excellent."
>
> Donald de Cogan, *International Journal of Numerical Modelling*

> "... instructive, applied and fun."
>
> David Jefferies, *The Times Higher Education Supplement*

> "This book integrates mathematical theory with a wide range of applications in physics ... it could be a stand-alone textbook for an upper-year undergraduate course, but would also provide a good accompaniment to undergraduate maths courses. Teachers may also be more generally interested in this book for its many examples that may serve as useful supplements to physics or geophysics courses ... I can highly recommend this book to students and teachers in the physical sciences."
>
> Mike Kendall, *Astronomy and Geophysics*

> "Guided by Roel Snieder's textbook, learning mathematical methods can be fun ... I highly recommend [it]."
>
> Francois Primeau, EoS

> "... with this preparation and a willingness to read with pen and paper in hand, any student is equipped to embark on a delightful journey."
>
> David Bressoud, *Journal of Undergraduate Mathematics and Its Applications*

225-AAZ-756

Front cover image: B-727 in flight during vortex study
wingtip smoke generators (Courtesy of NAS

Back cover image: Diagrammatic representation of a
wingtip vortex.

CAMBRIDGE
UNIVERSITY PRESS
www.cambridge.org

ISBN 0-521-83492-9

9 780521 834926